MOG 2007

Workshop on Multimodal Output Generation

CTIT Proceedings of the Workshop on
Multimodal Output Generation

Ielka van der Sluis, Mariët Theune, Ehud Reiter and Emiel Krahmer (eds.)

1495 University of Aberdeen

Tilburg University

University of Twente
Enschede - The Netherlands

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Sluis, van der. I., Theune, M., Reiter, E., Krahmer, E.

MOG 2007
Proceedings of the Workshop on Multimodal Output Generation
I. van der Sluis, M. Theune, E. Reiter, E. Krahmer (Eds.)
University of Aberdeen, United Kingdom
25-26 January 2007

ISSN 1574–0846
CTIT Workshop Proceedings Series WP07-01

trefwoorden: multimodality, natural language generation, multimodal generation,
modality choice, human modalities.

Book orders:
Ms. C. Bijron
University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
Human Media Interaction
P.O. Box 217
NL 7500 AE Enschede
tel: +31 53 4893680
fax: +31 53 4893503
Email: c.g.bijron@ewi.utwente.nl

Druk- en bindwerk: Reprografie U.T. Service Centrum, Enschede

Preface

It is a pleasure for us to welcome you in Aberdeen for the Workshop on Multimodal Output Generation (MOG 2007). Work on multimodal output generation so far has been mostly scattered across various events, so one of our objectives in organising MOG 2007 is to bring this work together in one workshop. Another objective is to bring researchers working in different fields together to establish common ground and identify future research needs in multimodal output generation. We believe the programme of MOG 2007 meets these objectives, as it presents a wide variety of work offering different perspectives on multimodal generation, while there is also the opportunity to meet colleagues, exchange ideas and explore possible collaboration.

We are very pleased to welcome three invited speakers. Jon Oberlander, from the University of Edinburgh in the UK, will give his personal view on multimodal output generation by embodied conversational agents. He argues that we need to carry out experiments tracking users' visual attention to investigate the impact of multimodal output on communication and task performance. Elisabeth André, from the University of Augsburg in Germany, will discuss corpus-based work that has been conducted in order to get insight on multimodal human-human dialogue with the aim to replicate such behaviours in an embodied conversational agent. Finally, Harry Bunt from Tilburg University in The Netherlands has kindly agreed to give us an update on the work of the ACL-SIGSEM Working Group on the Representation of Multimodal Semantic Information (see http://let.uvt.nl/research/ti/iso-tdg3/).

This volume brings together the papers presented at the MOG workshop together with abstracts provided by our invited speakers. In these workshop proceedings two different strands of work can be distinguished: half of the gathered papers present current work on embodied conversational agents (ECA's), while the other half presents current work on multimedia applications. Two general research questions are shared by all: what output modalities are most suitable in which situation, and how should different output modalities be combined? Below, the papers are briefly introduced.

To start with the work on ECA's, Adrian Bangerter and Eric Chevalley address the function of gestures in conversation by investigating when pointing gestures actually serve communication with respect to ambiguity and partner visibility. Their experiment provides evidence that either full or partial pointing gestures are used dependent of their communicative role. Mary Ellen Foster discusses the specific issues connected to multimodal corpora, different from unimodal corpora, that should be considered in corpus-driven generation systems. She advocates considering the annotation of contextual and cross-modal information and consideration of reproducibility of data in corpus building to allow for reuse of corpora and data-driven techniques. Markus Guhe proposes extensions of an incremental multimodal conceptualiser as a computational cognitive model of Levelt's conceptualiser. He argues for a late modal fission in a computational cognitive model of human language production in order to create a cognitively plausible multimodal dialogue model. Dirk Heylen discusses ongoing work on a sensitive artificial listening agent that tries to accomplish an attentive listening behaviour based on surface level cues. The process of data collection, analysis, and output evaluation used for this purpose illustrate that many different sources and methodologies need to be considered. Erwin Marsi and Ferdi van Rooden discuss the audiovisual expression of certainty and uncertainty in the context of question-answering systems, which is argued to be preferred over a verbal expression. They present a perception experiment that shows that certainty can reliably be expressed by a talking head that uses a limited repertoire of animated facial expressions. Paul Piwek challenges two assumptions: (1) non-verbal means of referring are secondary to verbal ones and, (2) speakers follow a single strategy to generate referring acts. He proposes two alternative strategies for modality selection based on correlation data obtained from an observational study that resulted in a corpus of task-oriented dialogues. Jan Peter de Ruiter gives an overview of some well-studied multimodal signals produced by humans and their implications for HCI. While distinguishing between functional and sensory modalities he addresses pointing gestures, eye gaze and spontaneous gestures, which are all meant to be interpreted by the listener, without being truly redundant.

With respect to the generation of multimedia presentations using modalities such as text and graphics, Yulia Bachvarova, Betsy van Dijk and Anton Nijholt propose a modality ontology that models the properties of different types of analogue and linguistic modalities and the relationships between them. Their goal is to model the kind of knowledge that can be used to automatically select the modality (or combination of modalities) that is most suitable to express a particular type of information, and to determine which modalities are most suitable to be combined. John Bateman and Renate Henschel present the main conclusions they have drawn from their empirical studies in the GeM project, in which they investigated the constraints imposed by multimodal genres. Their work focuses on the spatial-visual layout of multimodal documents, which they argue does not directly correspond to the rhetorical structure of the document. Charles Callaway discusses the multimodal presentation of non-localized, indoor route directions on a PDA. An experiment where subjects had to rely on different forms of information presentation to reorient themselves when lost in a building, provided useful lessons on how to improve the multimodal presentations. Christopher Habel and Cengiz Acartürk propose a multi-pass architecture for the automatic generation of multimodal documents. They argue that high-quality combinations of text and graphics can be generated through a revision process they call 'reciprocal improvement', which involves detecting possible gaps in co-reference between text and graphics. Charlotte van Hooijdonk and colleagues analysed a large corpus of multimodal presentations to find out which combinations of modalities should be used for the presentation of answers in a medical question-answering system. Their analysis revealed that visual media such as graphics, photos and animations significantly differ in terms of their function, and that the type of question (e.g., procedural or not) affects the choice of visual media. Željko Obrenović, Raphaël Troncy and Lynda Hardman propose a vocabulary of concepts that could be used to (semi-)automatically determine the accessibility of a particular multimodal interface. They describe how their vocabulary was defined by combining existing vocabularies of human functionalities and anatomical properties with a taxonomy of the interaction effects of various modalities. Finally, Yaji Sripada and Feng Guo describe a prototype system that generates multimodal presentations of dive computer data, consisting of a line graph with a textual summary. Evaluation of the system showed that the use of visual markers linking parts of the graph to their textual description was appreciated by the users, but it also revealed some remaining problems.

We believe that these proceedings present an excellent collection of the state of the art in multimodal output generation. Thanks are due to the programme committee members: John Bateman, Harry Bunt, Justine Cassell, Kees van Deemter, Betsy van Dijk, Roger Evans, Dirk Heylen, Fons Maes, Chris Mellish, Anton Nijholt, Rieks op den Akker, Richard Power, Graeme Ritchie and Marilyn Walker. Also thanks to our guest speakers and the authors of the submitted papers. At the University of Aberdeen and the University of Twente several people have helped us with the organization of this workshop. At the University of Aberdeen, Emma Stewart has done a terrific job in managing registration and accommodation. At the University of Twente, Hendri Hondorp provided invaluable assistance with preparing and LaTeX-editing the proceedings. Anja Annink-Tanke from CTIT managed the printing and publication of the proceedings.

MOG 2007 is endorsed by SIGGEN (ACL Special Interest Group on Generation) and has been made possible by financial support from the British Council and NWO (Netherlands Organization for Scientific Research) via the British Council - NWO Partnership Programme in Science. The workshop is also sponsored by NWO via IMOGEN (Interactive Multimodal Output Generation), a research project within the NWO-IMIX research programme. The research institute CTIT (Centre of Telematics and Information Technology) of the University of Twente kindly gave us permission to publish the proceedings of MOG 2007 in the CTIT Proceedings series. We are grateful to all these supporting organizations.

The organizers of this workshop,

Ielka van der Sluis, Mariët Theune, Ehud Reiter and Emiel Krahmer January 2007

MOG 2007 Programme Committee

Rieks op den Akker	University of Twente, The Netherlands
John Bateman	University of Bremen, Germany
Harry Bunt	Tilburg University, The Netherlands
Justine Cassell	Northwestern University, USA
Kees van Deemter	University of Aberdeen, UK
Betsy van Dijk	University of Twente, The Netherlands
Roger Evans	University of Brighton, UK
Dirk Heylen	University of Twente, The Netherlands
Emiel Krahmer	Tilburg University, The Netherlands (chair)
Fons Maes	Tilburg University, The Netherlands
Chris Mellish	University of Aberdeen, UK
Anton Nijholt	University of Twente, The Netherlands
Richard Power	Open University, UK
Ehud Reiter	University of Aberdeen, UK (chair)
Graeme Ritchie	University of Aberdeen, UK
Ielka van der Sluis	University of Aberdeen, UK (chair)
Mariët Theune	University of Twente, The Netherlands (chair)
Marilyn Walker	University of Sheffield, UK

Endorsed by SIGGEN

ACL Special Interest Group on Generation, http://www.siggen.org/

Sponsors

http://www.britishcouncil.org/

http://www.nwo.nl/

http://www.ctit.utwente.nl/

http://wwwhome.cs.utwente.nl/˜theune/IMOGEN/

Contents

Invited Speakers

Regular Speakers

What Are You Looking At?
A Personal View on Multimodal Output Generation

Jon Oberlander[*]
School of Informatics
University of Edinburgh
2, Buccleuch Place
Edinburgh, UK EH8 9LW
J.Oberlander@ed.ac.uk

Abstract

Embodied conversational agents (ECAs)—both virtual and real—can exploit multiple output modalities, including speech, facial expressions, head motions and hand gestures. Understanding the nuances of human expression is a research goal in itself. But not all of those nuances may matter to people interacting with an ECA. So, what is the balance between the benefits and costs of increasing the realism of ECAs' multimodal output? This paper draws on experience emerging from a series of studies to argue as follows. First, people may prefer greater realism, but it remains an empirical question as to whether realism improves communication. Secondly, even if it does, there is no guarantee that more expressive communication means better task performance. Thirdly, this may be because ECAs in complex visual tasks attract only minimal, intermittent attention. Thus, in current and planned work on human-robot interaction, we are studying the dual impact of multimodal communicative output on what people do, and on what they look at.

Keywords: Embodied conversational agents, human-robot interaction, eye-tracking.

1 Introduction

The EU IST JAST project includes work on human-robot joint action, which allows us to explore the role of multimodal communication where a human-robot pair constructs toy models together. The robot consists of a pair of mechanical arms with grippers, mounted in a position to resemble human arms, and an animatronic cat-like talking head capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The input channels include speech recognition, object recognition, face, gaze, and hand tracking, and force/torque sensors in the robot arms; the outputs include synthesised speech, head motions, facial expressions, and actions of the robot arms. Unsurprisingly, when considering how to engineer effective dialogues with this real, embodied conversational agent, we are inspired both by previous work on human-robot interaction (Sidner et al., 2005), and on previous work on virtual ECAs (Cassell et al., 2000; Ruttkay et al., 2006). But a recurring question which arises in either line of work is: what is the balance between the benefits and costs of increasing the realism of an ECA's multimodal output? This paper draws on personal experience emerging from a series of studies to argue that information about where people actually look will be of real value.

2 Preference for Naturalism vs Expressive Communication

First, people may prefer greater realism, but it remains an empirical question as to whether realism improves communication. Foster and Oberlander (2006) amongst others show that making facial

[*] This work has been completed in the context of the European Union Integrated Project JAST (Joint Action Science and Technology), grant FP6-IST2-003747. Special thanks to Mary Ellen Foster and to Sara Dalzel-Job.

motions more realistic (in a certain sense) makes people like them more. But this does not mean that the realism is making it easier for people to understand what an ECA is saying. Some of Foster's more recent, as yet unpublished results, also bear on this question, and specifically explore the extent to which redundancy across multimodal channels is detectable and useful.

3 Expressive Communication vs Task Performance

Secondly, however, even if greater realism helps communication, there is no guarantee that more expressive communication means better task performance. For instance, White et al. (2005) compare the effect on task performance of the expressiveness of a talking head. They suggest that some, but not all, users may actually be distracted by the talking head, thereby suppressing their task performance. This mirrors results on video-mediated communication, where it seems that having more multimodal information may actually lead to less effective task performance (O'Malley et al., 1996). And the task itself may have considerable impact, depending on whether or not it too demands visual attention. So, the effects of a talking head may be rather complex: to resolve them, we may have to look to information about on-line processing.

4 Attending (or not) to ECAs

It would be useful to find out what kinds of attention ECAs in complex visual tasks attract—and when. Unpublished pilot work by Masters students at Edinburgh, led by Dalzel-Job, bears on this question. Eyelink II eye-tracking technology was used to assess how and when users look at a gestural on-screen ECA in a route-following task. It is reported that whether or not the agent is gesturing, people fixate on it very infrequently (about 0.1% of the time), but that they do tend to look at it more frequently at the beginning and end of a sub-task.

5 Evaluating a Robot ECA

This leads us to our current plans for evaluating how humans interact with the JAST robot during collaborative construction tasks. We are studying the dual impact of multimodal communicative output on what people do, *and* on what they look at. Obviously, we are most interested in the effects of different dialogue strategies on task performance (including time, turns, and accuracy). But our robot can function with or without its cat-like head, and the head can be more or less expressive. Either way, the most functional version of the system incorporates tracking of the direction of the human partner's gaze. Given this, we hope to establish when—and ultimately why—multimodal output from a robot ECA either attracts or guides attention, and when this actually helps their partner get things done.

References

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (2000). *Embodied Conversational Agents.* MIT Press.

Foster, M. E. and Oberlander, J. (2006). Data-driven generation of emphatic facial displays. In *Proceedings of EACL 2006*, pages 353–363. Trento, Italy.

O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., and Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction Interacting with Computers, **8**, 177-192

Ruttkay, Z., André, E., Johnson, W. L., and Pelachaud, C. (2006). Evaluating Embodied Conversational Agents. In Ruttkay, Z., André, E., Johnson, W. L., and Pelachaud, C., editors, *Evaluating Embodied Conversational Agents*, number 04121 in Dagstuhl Seminar Proceedings.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1–2):140–164.

White, M., Foster, M.E., Oberlander, J., and Brown, A. (2005). Using facial feedback to enhance turn-taking in a multimodal dialogue system. In *Proceedings of HCI International 2005.*

From Annotated Multimodal Corpora to Simulated Human-like Behaviors

Elisabeth André
Lehrstuhl für Multimedia-Konzepte und Anwendungen, Institut für Informatik
Universität Augsburg, Eichleitnerstr. 30, D-86159 Augsburg Germany

Abstract

A number of approaches to modeling the behaviors of embodied conversational agents (ECA's) are based on a direct simulation of human behaviors. Consequently, it comes as no surprise that the use of data-driven approaches which allow us to validate design choices empirically has become increasingly popular in the ECA field. To get insight into human-human conversation, researchers rely on a large variety of resources including recordings of users in "natural" or staged situations, TV interviews, Wizard of Oz studies, and motion capturing data. Various annotation schemes have been designed to extract relevant information for multimodal behaviors, such as facial expressions, gestures, postures and gaze. In addition, there has been increasing interest in the design of annotation schemes to capture emotional behaviors in human-human conversation. Progress in the field has been boosted by the availability of new tools that facilitate the acquisition and annotation of corpora.

The use of data-driven approaches provides a promising approach to the modeling of ECA behaviors since it allows us to validate design choices empirically. Nevertheless, the creation of implementable models still leaves many research issues open. One difficulty lies in the fact that an enormous amount of data is needed to derive regularities from concrete instantiations of human-human behavior. In rare cases, we are interested in the replication of behaviors shown by individuals. Rather, we aim at the extraction of behavior profiles that are characteristic of a group of people, for example, introverts versus extroverts. Furthermore, the resulting ECA behaviors only emulate a limited amount of phenomena of human-human behaviors. In particular, the dynamics of multimodal behaviors has been largely neglected so far. Last but not least, there is the danger that humans expect a different behavior from an ECA than from a human conversational partner which might limit the potential benefits of a simulation-based approach.

In my talk, I will provide an overview of existing corpus-based work that has been conducted in order to get insight on multimodal human-human dialogue with the aim to replicate such behaviors in an embodied conversational agent. I will present several approaches to bridge the gap from corpus analysis to behavior generation including copy-synthesis, generate-and-filter as well as first attempts to realize trainable generation approaches. Finally, I will discuss several empirical studies that have been conducted with the aim to validate models derived from a corpus.

Keywords: Embodied Conversational Agents, Multimodal Corpora

Towards a Unified Knowledge-Based Approach to Modality Choice

Yulia Bachvarova, Betsy van Dijk, Anton Nijholt
Human Media Interaction Group, University of Twente,
PO BOX 217, 7500 AE Enschede, The Netherlands
{y.s.bachvarova, e.m.a.g.vandijk, anijholt}@cs.utwente.nl

Abstract

This paper advances a unified knowledge-based approach to the process of choosing the most appropriate modality or combination of modalities in multimodal output generation. We propose a Modality Ontology (MO) that models the knowledge needed to support the two most fundamental processes determining modality choice – modality allocation (choosing the modality or set of modalities that can best support a particular type of information) and modality combination (selecting an optimal final combination of modalities). In the proposed ontology we model the main levels which collectively determine the characteristics of each modality and the specific relationships between different modalities that are important for multi-modal meaning making. This ontology aims to support the automatic selection of modalities and combinations of modalities that are suitable to convey the meaning of the intended message.

Keywords: Modality Ontology, Modality Choice, Modality Allocation, Modality Combination.

1 Introduction

The process of choosing and combining modalities to best convey the intended message is central for multimodal output generation. It is also a complex and highly knowledge-intensive process that depends on the type of the information that has to be represented and the specifics of the context, the user and the particular goal of the multimodal presentation on the one hand and the proper understanding and modelling of the nature of each modality and of multimodal meaning making on the other hand. Research on all these different aspects has been conducted by different communities. A lot of the research results gained, though relevant for multimodal output generation, remain scattered and not really employed to their potential. A unified framework, capturing the aforementioned aspects in their array of dependencies can properly address and formalize the complexity of the problem of modality choice.

The work that we present in this paper attempts to start addressing the issues related to modality choice in a unified and systematized manner. The two most fundamental processes related to modality choice are modality allocation and modality combination. Modality allocation assigns the most appropriate modalities that can best represent the types of information that have to be represented. Modality combination is the process where modalities are integrated into a coherent final multimodal message.

We start with the assumption that there is a formal representation, for example a domain ontology, of what has to be represented. We look at the types of information that have to be represented and the existing relationships between them and map this to the specific features of modalities describing their strengths and weaknesses in representing such information types and relationships. We further apply principles for optimal cognitive information processing or exploit the interdependencies between different modalities that determine multimodal meaning making in order to generate the most optimal modality combination(s).

Central in the design of the ontology is the idea that there are two main aspects that properly describe each modality – the content it represents, and its nature. While modelling the content of some modalities has recently received significant attention, research on the nature of a modality has not been properly systematized. Therefore, we address the issue of what describes the nature of a modality void of its content. Moreover, the focus on the relation between modality content and modality form will be shown to have important implications for multimodal meaning making.

We start by describing the main levels of the Modality Ontology providing examples on how the knowledge modelled in these levels can support modality choice. We then provide an example of the relation between modality content and modality profile. Finally, we conclude by outlining our future research directions.

2 MODALITY ONTOLOGY

The main purpose of the ontology we propose is to be able to support the automatic selection of modalities and combination(s) of modalities, hence the processes of modality allocation and modality combination. To be able to support these two processes, the Modality Ontology (MO) has to model the following main types of knowledge about modalities - knowledge about the capacity of each modality to represent different types of information, knowledge about the cognitive and perception related aspects of each modality's nature, and knowledge about the structural dependencies that exist between the different modalities and that determine the syntax of a given modality combination.

We demonstrate, but not in detail, how each of these aspects of knowledge about modalities is modelled by the ontology and provide simple examples how the ontology can support modality allocation and combination.

2.1 THE UPPER LEVEL OF THE MODALITY ONTOLOGY

The central idea of the approach we advance in this paper is that the meaning that each modality carries is determined by its content (the particular information it represents), its nature per se, that is its content-independent characteristics, and the relations existing between these two main aspects. In the MO the nature of a modality is modelled by the profile level. Further in this subsection we describe this level in more detail. Figure 1 shows the upper level of MO where the Modality class represents the operational concept of the ontology. Modality presents ModalityContent and is described by ModalityProfile.

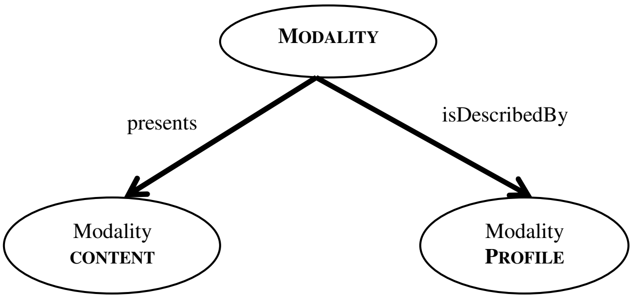


Figure 1: Upper level of the Modality Ontology

The ModalityProfile class describes knowledge about modalities at three different main levels – the information presentation level, the perception level, and the structural level. In MO these three levels are presented by the classes InformatonPresentationProfile, PerceptionProfile and StructuralProfile respectively (see Figure 2).

2.1.1 Information presentation level

The *information presentation level* models those modality characteristics that describe the strengths and weaknesses of each modality in representing particular types of information.. At the upper level of the InformationPresentationProfile we distinguish between *linguistic* and *analogue* modalities. The characteristics *linguistic* and *analogue* have been chosen based on their argued generality and robustness in profoundly distinguishing the different capabilities of modalities in representing information (Bernsen, 1994; Stockl, 2004). Linguistic representations, such as text and speech, are based on existing syntactic-semantic-pragmatic systems of meaning (Bernsen, 1994). An important feature of linguistic representations is that they *lack specificity* (Stenning & Oberlander, 1991); that is, they cannot specify precisely how things, situations or events look, sound, feel, smell or taste. Instead, linguistic representations are *abstract* and *focused* – they focus at some level of abstraction on the subject matter to be communicated. Those characteristics of linguistic representations determine their strength in representing abstract concepts, states of affairs and relationships. Analogue representations, such as images, represent through aspects of similarity between the representation and what they represent (Bernsen, 1994; Stockl, 2004). This determines the strong capacity of analogue representations to portray essentially visual or spatial information (Tversky, Morrison & Betrancourt, 2002). Analogue representations lack focus and can only to a limited extent represent abstract information. Knowing which modality feature is responsible for representing which information type allows mapping between what has to be represented and the modalities which can actually do that, i.e., MO supports the automatization of the modality allocation process. The information presentation level of the ontology can also support the modality combination process. The features of linguistic and analogue modalities we have chosen to describe here are complementary. The complementarity of features of analogue and linguistic modalities determines their frequent use together. In Section 2.1.2 we provide a concrete example of how modality combinations based on complementarity can be calculated.

The features of analogue and linguistic modalities that determine their capacity in representing different types of information are members of the class AnalogueModalityFeatures and LinguisticModalityFeatures respectively (see Figure 2).

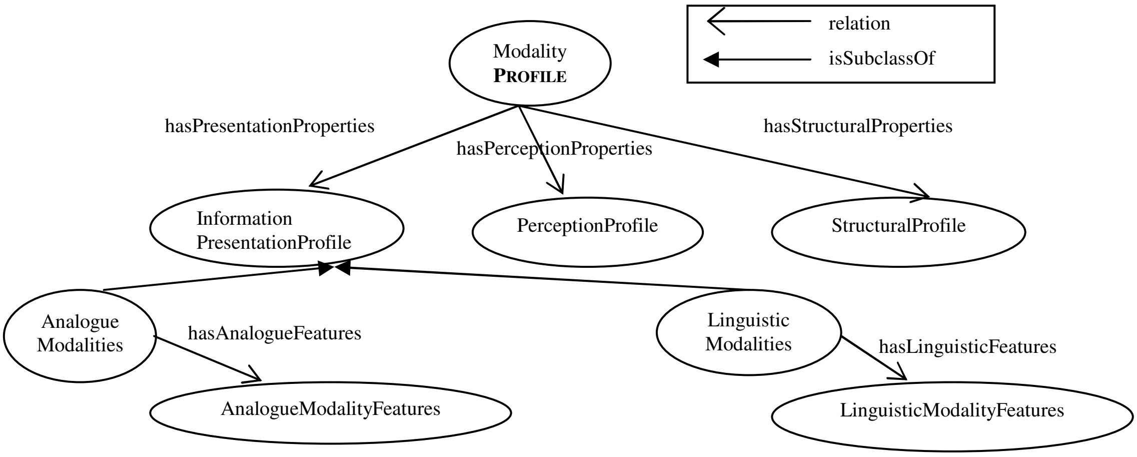


Figure 2: Information presentation level of the Modality Ontology

2.1.2 Perception level

The perception level models those modality characteristics which determine how a particular modality is perceived and processed by the human perceptual-sensory system (see Figure 3). At this level we distinguish between visual, auditory and haptic modalities. Visual are the modalities that are perceived through the visual sensory channel, for example written text or images. In the ontology visual modalities are represented by the class VisualModalities. Auditory modalities are perceived through the auditory sensory channel, for example speech or music and are represented in the ontology by the AuditoryModalities class. Haptic modalities are related to the sensory system of touch. This modality class falls out of the scope of interest of this paper.

An important dimension in the way a particular modality is processed is the time allowed for its processing. Static modalities, for example pictures or static text, allow unlimited time for inspection and processing. In contrast, dynamic modalities (animation, video) are transient and do not allow freedom of perceptual inspection. In MO static modalities are represented by the class StaticModalities and dynamic modalities are represented by the class DynamicModalities.

We further describe an example of how the knowledge modelled by the perception profile can support the process of modality combination by generating multimodal output in accordance with well established principles for cognitive information processing. More concretely, our example demonstrates how to generate multimodal combinations that comply with the cognitive Modality Principle postulated and empirically tested in (Moreno & Mayer, 1999). This principle states that when giving multimedia explanations words should be presented as auditory narration rather than as visual on-screen text. The Modality Principle is based on two important themes from theories of human cognitive processing (Baddley 1992; Chandler & Sweller, 1991; Pavio 1986): (i) the processing capacity (or working memory capacity) of the visual and auditory information-processing channels is limited and (ii) active processing involves selecting relevant visual and verbal information, organizing the material into coherent mental models and integrating between visual and verbal representations as well as existing knowledge from the long-term memory. In accordance with (ii) the combination between visual and verbal information (in MO between linguistic and analogue modalities) is realized based on the complementarity of the features specificity, abstractness and focus (see description of information presentation level). Avoiding cognitive overload (i) will require that the above generated combination is also a combination between visual and auditory modalities (in MO modelled by the perception level). Thus the final combination is calculated to be between a modality belonging to the linguistic and auditory classes, that is, speech, together with analogue visual modality, for example animation. The choice of which analogue modality to use can be subject to applying additional principles or design rules. Generating multimodal output based on the modality cognitive principle makes use jointly of the information presentation and perception levels of the proposed Modality Ontology by applying modality combination rules.

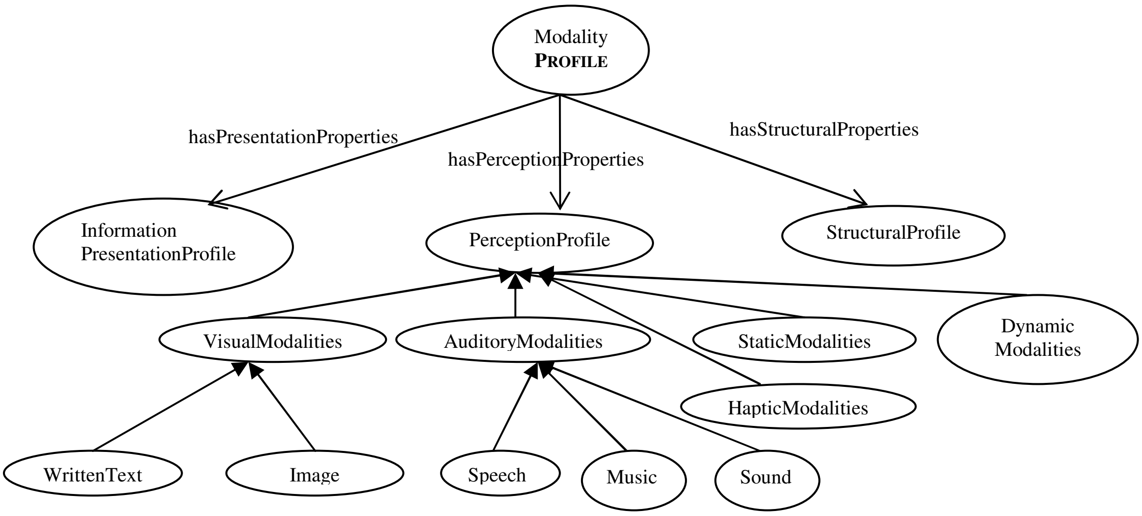


Figure 3: Perception level of the Modality Ontology

2.1.3 Structural level

The structural level models the structural dependencies that can exist between the composite modalities of a multimodal presentation. Structural dependencies form the syntax of multimodal presentations and as such have a direct bearing on the way multimodal messages construct and convey meaning. For an illustration consider the structural dependence of a substrate (background) and the information carried by the modality situated on that substrate. By virtue of being a substrate one modality can determine the interpretation scope and provide the semantic context of the modality which is situated within the substrate. A more concrete example is the combination of an icon on a map substrate. The map used to describe a region of the world possesses an internal structure – points on it correspond to points in the region it charts. When used as a background of an icon, one may indicate the location of the object represented by the icon by placing it in the corresponding location on the map substrate (Arens, Hovy & Vossers, 1993).

Pertinent to the structural level is the distinction between dependent and independent modalities made by (Bernsen, 1994). Independent modalities can do much of their representational work on their own; for example text alone can express almost everything. In contrast, dependent modalities need other modalities to serve representational purposes. Graphs are examples of dependent modalities as they almost always require clear and detailed linguistic annotation for their interpretation. Structural dependencies are important for calculating modality combinations. We have chosen to model these dependencies as properties relating the classes of modalities forming the dependency and not necessarily as part of the structural level. For example, in the ontology the classes Graphs and Labels (see Section 2.2.2 for more in-depth explanation) are related by the inverse properties *annotates* and *areAnnotatedBy*.

2.2 ANALOGUE AND LINGUISTIC MODALITIES

At this level of the ontology we describe which more specific differentiations can be made between modalities in terms of their capacity to represent different types of information. The members of each modality class at this level are characterized not only by the set of features related to that particular level by also, through inheritance, by the set of features characterizing the upper level.

2.2.1 Analogue modalities

Zooming in on the AnalogueModalities class, it comprises of the disjoint classes of Images, Maps, Graphs and Diagrams (see Figure 4). This classification is based on Bernsen's taxonomy of output modalities and Lohse's classification of visual representations (Lohse et al., 1994). The specific characteristics describing the way each of these modalities represents information are members of the classes ImageFeatures, MapFeatures, GraphFeatures and DiagramFeatures respectively. Table 1 presents some of the features characterizing images, graphs, maps and the three types of diagrams – structural, process and conceptual. The features have been selected from existing literature describing the different characteristics and aspects related to the nature of different modalities (Bernsen, 1994; Lohse et al., 1994; Tufte, 1983; Twyman, 1979). This set of features is by no means exhaustive. It is not the aim of this paper to describe such exhaustive set, but just to illustrate the approach we adopt in modeling the knowledge about modalities.

MO represents modalities at levels deeper than the specific attention of this subsection. For example, graphs can be scatterplot, categorical, line, stacked bar, bar, pie, box, fan, response surface, histogram etc. Each of these graph types has specific characteristics which distinguish it as a type on its own. In a fashion similar to the one applied for the aforementioned ontology levels, each graph type is a modality class which is related to the class of properties describing this modality.

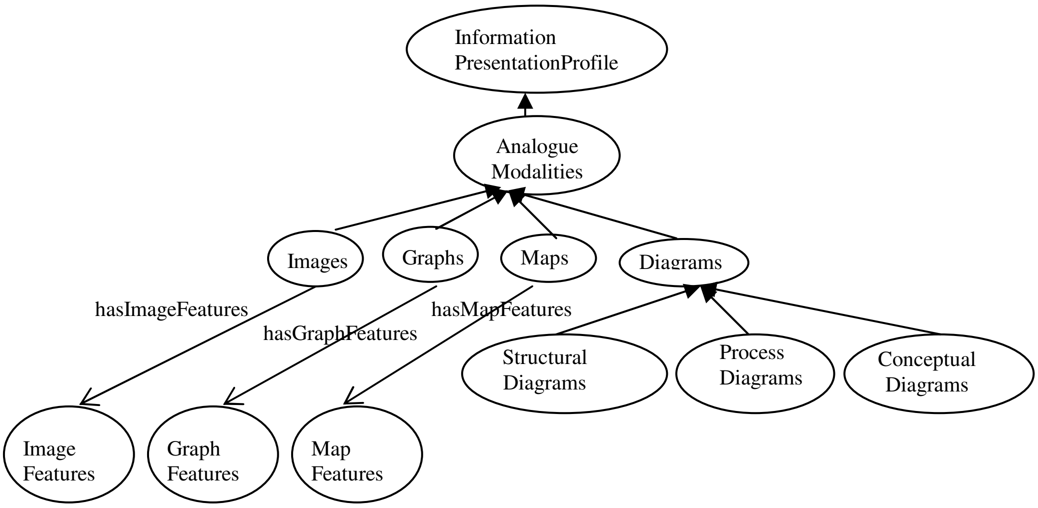


Figure 4: Analogue modalities

Modality	Information presentation related features
Image	- high specificity - full correspondence with the represented object - preserves distance properties of real world space - preserves interval properties of real world space
Map	- represents physical geography - represents location - represents relational structure of objects and events
Graph	- encodes quantitative information - emphasizes the whole display - symbolic (no recognizable similarity to the subject matter or domain of representation) - supports analysis of data information - supports reasoning about data information
Conceptual Diagram	- presents analytical decomposition of an abstract entity - facilitates the perception of structure and relationship
Structural Diagram	- describes a physical object - conveys spatial, nonnumeric, concrete information
Process Diagram	- describes the interrelationship and processes associated with physical objects - the spatial data expresses dynamic, continuous or temporal relationships among the objects

Table 1: Information presentation related features of analogue modalities

2.2.2 Linguistic modalities

At the *linguistic modalities* level the main distinction is between *text, discourse, label* and *notation linguistic modalities* (see Figure 5). The distinction between *text* and *discourse* modalities stems from the different behaviour of written language and spontaneous spoken language. While written language is situation independent, i.e., the recipient and the author of the communication do not need to share the same space, time and situation, spoken language has evolved to serve situated communication. Label and notation modalities are brief expressions of focused information. These features make labels well suited in combinations with modalities that require short textual annotation, for example graphs or conceptual diagrams. Relationships of that kind are directly encoded in the ontology (see Figure 5) and can be used for a straightforward calculation of certain modality combination. In the particular example with graph and label modalities the properties *annotate* and *areAnnotatedBy* are inverse. It is possible to specify that in OWL using owl:inverseOf.[1] *Notations* are for specialist users and their most prominent feature is limited expressiveness.

Similarly to the depicted relations between modalities and their features at the different already described levels of the ontology, all the classes of linguistic modalities are described by their corresponding features. We did not choose to show all the feature classes of linguistic modalities in Figure 5 as our attention is mainly on depicting the new aspects of knowledge about modalities that each level introduces.

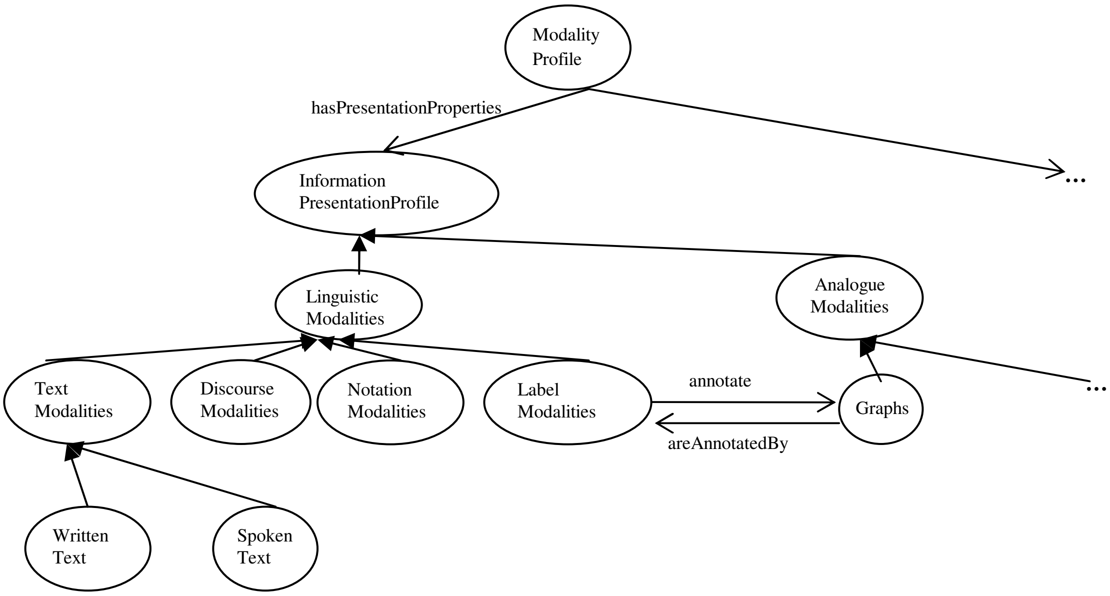


Figure 5: Linguistic Modalities

2.2.3 Information channels

Information channels are an important aspect determining the way modalities convey information. Information channel has been defined as a perceptual aspect (an aspect accessible through human perception) of some medium which can be used to carry information in context (Bernsen, 1994) or an independent dimension of variation of a particular information carrier in a particular substrate (Arens, Hovy & Vossers, 1993). An example of the latter definition would be an icon that can convey information by its shape, color and position and orientation in relation to a substrate map. What Bernsen (2004) and

[1] MO has been implemented in OWL.

Arens et al., (2003) call information channels, Stockl (2004) calls sub-modes, defining them as the building blocks of a core mode's grammar (core modes correspond to the level describing linguistic and analogue modalities in MO). In what follows we describe the way the information channels of typography and colour are modelled by MO. The approach applied for these two information channels can be generalized for the remaining information channels.

Typography is an important aspect in representing written text and can contribute to its meaning beyond the linguistics. We have chosen to model typography at the profile level (see Figure 6) because it is related to the modality form, i.e., one and the same typography can accommodate different contents. In the following subsection we will use this ontological distinction to demonstrate how MO can capture important meaning making relations between content and profile. The class Typography contains all the main constructs that describe typography, such as font type and size, spacing, paragraphing, margins, etc. In the ontology they are modelled as subclasses (Paragraphing, Font, Colour) of the class Typography.

Colour is an information channel that describes not only typography but also images. In order to properly capture all the important features that describe colour we align MO at this level with the MPEG-7 ontology (Hunter, 2005) and more specifically with the MPEG-7 colour visual descriptor. In the MPEG-7 ontology some widely used visual and audio features or properties are represented by a choice of descriptors. The visual properties described are colour, texture, shape and motion while the audio properties are silence, timbre, speech, musical structure and sound effects. The property colour is described by the descriptors (classes in the MPEG-7 ontology) DominantColour, ScalableColour, ColourLayout, ColourStructure, GoFGoPColour (see Figure 6).

We follow the same approach of aligning with MPEG-7 ontology feature descriptors when describing the remaining information channels.

3 RELATIONSHIP BETWEEN CONTENT AND PROFILE

To model the relationship between content and profile we need a proper representation of modality content in addition to the modality profile representation we describe in this paper. Modeling content is not our focus and for that reason we try to make use of already existing frameworks. At the content level we align with the MPEG-7 ontology and more specifically the part that concerns content representation (see Figure 7).

To illustrate the capacity of MO to capture and model meaning that is derived from the relationship between modality content and profile we use an example described in (Stockl, 2004). We have chosen this example because of the necessity it poses on modelling content and profile separately and establishing a connection between the two.

The example is that of an advertisement of the RSPCA (the Royal Society for the Prevention of Cruelty to Animals) for free range eggs where the verbal text is typographically designed to yield the visual form and appearance of a supermarket receipt. The language contained in the receipt is not what we would normally expect to read on a receipt (the bought items and their prices) but the textual message of the advertisement (the appeal to people not to buy battery eggs). In this example the exported typographical repertoire has a semantic impact. The receipt form of the text makes the pivotal point that it is in the supermarket where farming policies are shaped via the price of the eggs and consumer behaviour.

In order to be able to capture or generate such sophisticated interplay between content and form we need to have the proper frameworks to model the two aspects separately as well as their relations.

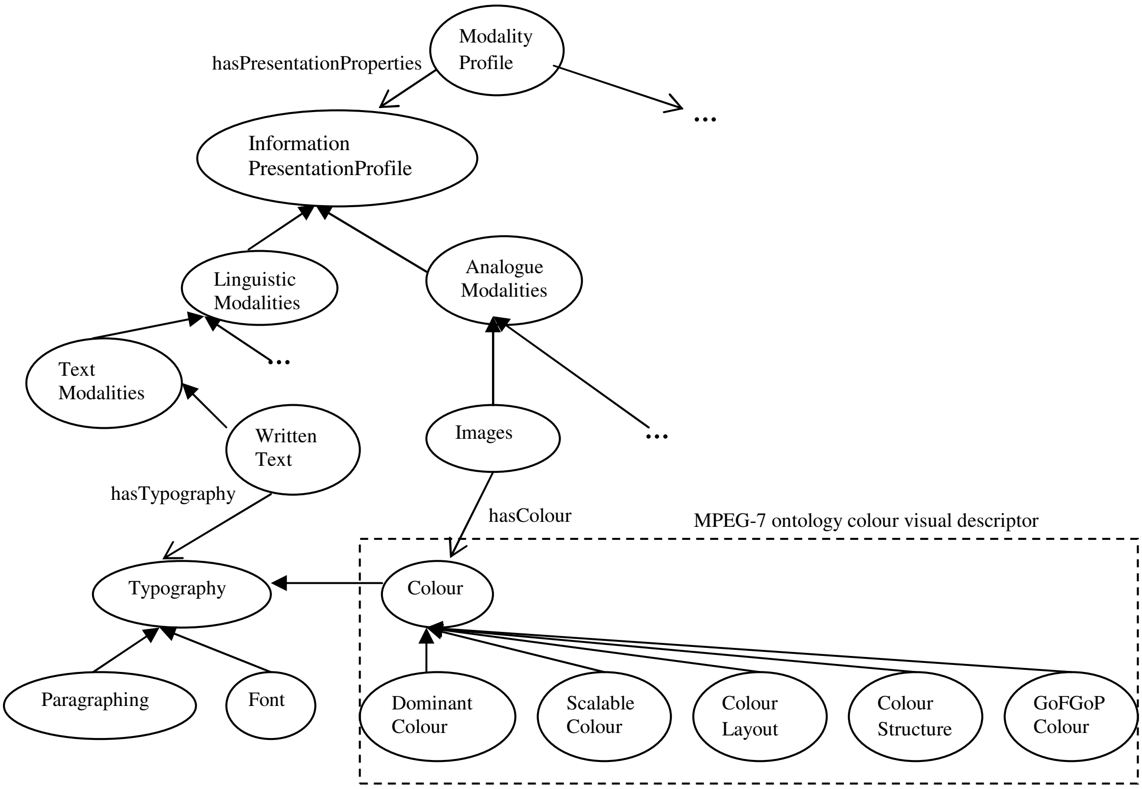


Figure 6: Information channels

Using MO the representation provided by the RSPCA advertisement can be modelled on the content and profile levels. On the profile level we describe the specific features of the typography of a supermarket receipt (the specific type and size of the font, paragraphing, etc.) and relate it to the concept of a supermarket receipt. On the content level the representation of the pair - item and its corresponding price - is also related to the concept of supermarket receipt as this is the information that you normally find on a receipt. The instantiations of the specific content of a receipt and its typography are related by the hasTypography relation. In other words, text which says which items have been bought and in what price has a specific typography – narrow margins marked by lines of three stars each, dotted font typical for cash-desk printer, etc. The text and the typography are both characteristic for supermarket receipts. Now when in this relationship between content and form only the content is changed, in our particular example the content of the receipt is substituted with the advertisement text, the advertisement text appears in the form of a supermarket receipt (the hasTypography relation to the receipt typography stays unchanged). The meaning derived from the new representation is a combination of the meaning derived from the content level (what the text of the advertisement says) and the meaning associated with the specific instantiation of the Typography class, that is a supermarket receipt. In other words content and form (profile) shift and blend and users translate or transpose meaning from one of those two aspects to the other.

.

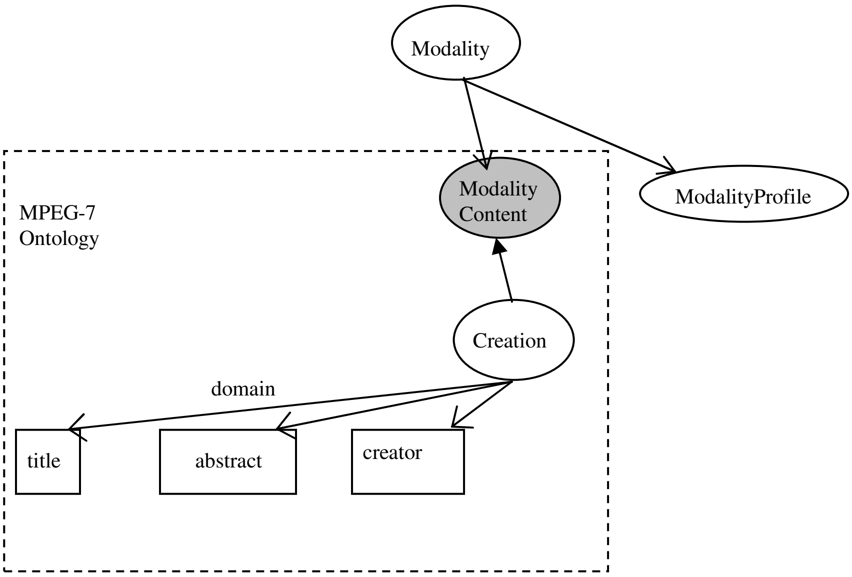


Figure 7: Aligning with the MPEG-7 ontology on the content level

4 CONCLUSION AND FUTURE WORK

The processes of modality allocation and modality combination are knowledge intensive and require proper representation of the knowledge that supports them. The Modality Ontology we propose models that knowledge at three different levels – properties of modalities that determine their capacities to represent different types of information, properties that determine the way each modality is perceived and processed by human cognitive systems and structural dependencies between different modalities. The knowledge described on the first level supports mainly the modality allocation process while the second and the third level are used for calculating modality combinations. MO has the capacity to serve as a unified framework that captures different aspects of knowledge about modalities that have already been modelled for different purposes by different research communities. We have demonstrated a possible alignment with the MPEG-7 ontology.

We are currently developing more robust and generalized methods for modality allocation.

ACKNOWLEDGEMENTS

This work was supported by the ICIS program (http://www.decis.nl/html/icis.html). ICIS is sponsored by the Dutch government under contract BSIK 03024.

REFERENCES

Arens, Y., Hovy, E. & Vossers, M. (1993) On the knowledge underlying multimedia presentations. In *Intelligent Multimedia Interfaces.* Mark Maybury, editor. AAAI Press.

Baddeley, A. (1992). Working memory. *Science, 255.*

Bernsen, N. O. (1994). Foundations of multimodal representations: A taxonomy of representational modalities. *Interacting with Computers*, 6(4):347–371.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*.

Hunter, J. (2005). Adding multimedia to the Semantic Web - Building and applying an MPEG-7 ontology. *Multimedia Content and the Semantic Web: Standards, Methods and Tools*, Giorgos Stamou and Stefanos Kollias (Editors), Wiley.

Lohse, G., Biolsi, K., Walker, N., & Rueter, H. (1994). A classification of visual representations. *Communications of the ACM*, v.37.

Moreno, R. & Mayer, R.E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, 91, 358-368.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. Oxford, England: Oxford University Press.

Stenning, K. & Oberlander, J. (1991). Reasoning with words, pictures and calculi: Computation versus justification. In Barwise, J., Gawron, J.M., Plotkin, G., and Tutiya, S. (Eds.). *Situation Theory and Its Applications.* Stanford, CA:CSLI, Vol.2.

Stockl, H. (2004). In between modes. Language and image in printed media. In E.Ventola, C. Charles, and M. Kaltenbacher (Eds.). *Perspectives on Multimodality*. John Benjamins Publishing Company.

Tversky, B., Morrison, J.B., & Betrancourt, M. (2002). Animation: Can it facilitate? *Int. J. Hum.-Comput. Stud.* 57, 4.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut.

Twyman, M. (1979). A schema for the study of graphic language. In Kolers, P., Wrolstad, M., and Bourna, H. (Eds.). *Processing of Visible Language* Vol. 1. New York: Plenum Press.

Pointing and Describing in Referential Communication: When Are Pointing Gestures Used to Communicate?[1]

Adrian Bangerter and Eric Chevalley
Institut de Psychologie du Travail et des Organisations
University of Neuchâtel
Rue de la Maladière 23
CH-2000 Neuchâtel, Switzerland
adrian.bangerter@unine.ch

Abstract

What functions, if any, do gestures serve in conversation? Recently, it has been suggested that the function of gesture may vary within a particular type, as a function of the immediate communicative context. Thus, it may not be a question of *whether* a type of gesture is used to communicate, but *when*. We investigated this possibility for the case of pointing gestures in a referential communication task. Pairs worked together to identify targets (photos of people) among pictures in an array that they both could see. They were free to talk and gesture. We manipulated two factors: ambiguity (the number of pictures in an array) and partner visibility. Visible pairs could see each other, hidden pairs could not. We distinguished between full and partial pointing gestures. Full points involve raising the arm whereas partial points do not. Full points were used more often by visible than hidden pairs, thereby suggesting that they are intended to communicate. Their use reduced verbal effort. Their frequency decreased with increasing ambiguity. The use of partial points did not vary with visibility, suggesting that they are automatic in production.

Keywords: Pointing gestures, conversation, referential communication.

1 INTRODUCTION

1.1 THE FUNCTION(S) OF GESTURES IN CONVERSATION

What functions do gestures serve in conversation? Given their ubiquity, one seemingly obvious function is communicating. By this view, gestures are used by the speaker to add information to speech, e.g., by expressing semantic content complementary to verbal utterances. However, a long and diverse strand of research has questioned this possibility. In an influential article, Rimé and Schiaratura (1991) proposed that co-speech gestures are primarily related to speech production, and that their communicative function is peripheral. Using a wide range of methods and research strategies, experimental studies have indeed shown that gestures are related to speech production (Krauss, 1998). One study by Chawla and Krauss (1994) found that gestures are more frequent in spontaneous speech, where production is more demanding, than in rehearsed speech. Another (Morrel-Samuels & Krauss, 1992) demonstrated that gestures concurrent with unfamiliar words are longer in duration than those concurrent with familiar words. Gestures also more generally aid in organizing cognition, as when they facilitate counting activities

[1] This study was supported by a grant (101411-103472) from the Swiss National Science Foundation to the first author.

in preschoolers (Alibali & DiRusso, 1999) or maintain information in spatial working memory (Morsella & Krauss, 2004).

According to these findings, gestures facilitate lexical retrieval, message conceptualization (Kita, 2000), or thinking, in other words they are functional for speakers. To demonstrate a communicative function of gestures in conversation, it is necessary to show (1) that they are designed by speakers for addressees, and (2) that they actually have an impact on comprehension (Krauss, Morrel-Samuels, and Colasante, 1991), in other words the demonstration must focus on activities of both speakers and addressees. Moreover, gestures should be elicited in communicative rather than non-social situations (Alibali, Heath, and Myers, 2001). Indeed, researchers arguing for a communicative function of gesture (Bavelas, 1994; Kendon, 1994) often point out the importance of studying spontaneous gestures produced in natural conversation and their relation to the immediate communicative context.

The possibility that gestures are designed by speakers for addressees constitutes a case of audience design (Clark & Murphy, 1982). Audience design can be shown by comparing the frequency of gestures produced in situations where communicators are visible to each other with the frequency of gestures produced in situations where they are not. If gestures are designed by speakers for addressees, then a lack of visibility should suppress gesture production. Note that, even in this case, the question of the precise level of intentionality of the gesture remains difficult to answer (see however Melinger & Levelt, 2004). If, however, gesture production is insensitive to this change in context, then that is a strong indicator that it is automatic. Gesture frequency does indeed differ according to visibility (Alibali et al., 2001), but not systematically (Rimé, 1982). Sometimes, gestures are produced more frequently in visible conditions but are not entirely suppressed when communicators are not visible to each other. A particularly subtle variation of this technique was used by Özyürek (2002). She found that speakers changed the orientation of their gestures to accommodate to addressees seated at different angles.

Showing that gestures facilitate comprehension is more difficult. Sometimes, gestures are not attended to by addressees, are not well-remembered, or are not easily interpreted independently of co-occurring speech (e.g., Krauss, Morrel-Samuels, & Colasante, 1991; Krauss, Dushay, Chen, & Rauscher, 1995). Early studies (e.g., Graham & Argyle, 1975) have, however, found that gestures can facilitate comprehension in communicative situations. Gestures can even enhance the impact of television advertisements (Beattie & Shovelton, 2005). Other work (Bangerter, 2004; Kelly, Barr, Church, & Lynch, 1999; Louwerse & Bangerter, 2005; Thompson & Massaro, 1986) has shown that pointing gestures affect comprehension.

Of course, the possibility that gestures are related to speech production is not incompatible with the possibility that they are communicative (Alibali et al., 2001; Bavelas, 1994; Özyürek, 2002). These functions are interrelated in other aspects of language use. For example, fillers (e.g., *uh* or *um*) reflect both speech production and interactional processes (Clark & Fox Tree, 2002). Thus, the finding that gestures facilitate speaking does not constitute evidence against the claim that they are communicative, especially when they have been elicited in non-communicative situations (Melinger & Levelt, 2004). Gestures may have both speaker-related (i.e., speech production) and addressee-related (i.e., communicative) functions. In the wake of this conclusion, an emerging trend in the field is the detailed study of different types of gestures and their relation to speech and to the immediate conversational context. Gestures are increasingly viewed as multifunctional and integrated with speech (e.g., Bavelas & Chovil, 2000). Alibali et al. (2001) found that the production of representational gestures (a category subsuming iconic, deictic and metaphoric gestures) varied according to visibility but that the production of beat gestures did not. Nonetheless, representational gestures were also produced in non-visible conditions, indicating that they may serve both cognitive and communicative functions. The authors suggested that research on gesture should "examine how different speakers use gestures in different types of contexts for both speaker-internal and communicative purposes" (p. 186). Bavelas (1994) even goes beyond this in arguing against a purely taxonomic approach to gesture: "In short, the goal of analysis should not be to decide in which category we should put a gesture (or all gestures) but rather to discover at least some of the things a gesture is doing at its particular moment in the conversation." (p. 204).

In sum, it has been proposed that gestures be considered as multifunctional, in other words that different types of gestures be compared, that functional differences may hold even within gesture types, and therefore that they be studied in detail and in relation to their conversational context. Taking into account the immediate context of gesture production may reveal *when* they are communicative and when they are not, rather than *whether* gesture in general is communicative or not or whether a particular type of gesture is communicative or not. The present study takes up these calls. We investigated pointing gestures in a referential communication task. Our goal was to investigate the precise functions of pointing, be they communicative or not. We conducted a controlled experimental study that nonetheless allowed for conversational interaction. Such a setting allows manipulation of key variables (e.g., mutual visibility of interaction partners) while retaining features of naturalistic situations to analyze in detail the relation between gesture and speech. The case of pointing is theoretically interesting because it has received comparatively less attention than other types (for example iconic gestures), and also because pointing seems to convey different information or have different functions than, say, iconic gestures. In the next section we briefly review research on pointing gestures, focusing on the relation between pointing and speech in conversational situations between adults.

1.2 POINTING GESTURES

Pointing gestures are a particular form of deixis. Classical theorists have described their relationship to other types of gestures and to language. Peirce (see Buchler, 1940) classified signs into three categories: icons, indices and symbols. Icons bear a perceptual resemblance to their object of reference (e.g., iconic gestures that mimic a particular manner of movement). Symbols (such as words) are arbitrarily related to their referents. Indices (e.g., pointing gestures) are related to their referents by a physical connection. They are ways of focusing attention. More recently, Clark (2003) proposed that there are two basic kinds of indices, pointing and placing. With pointing, people move the gaze of their addressee towards the referent. With placing, people move the referent into the field of view of the addressee. Pointing and placing thus constitute two ways to focus attention. Another early theorist, Bühler (1965), argued that deictic words primarily serve to direct the attention of the addressee. According to the functional view of gesture advocated by Bavelas (1994), different kinds of gestures (e.g., manual pointing or gaze) can be considered deictic insofar as they serve to focus visual attention. It is even possible to conceive of "chains" of indices. In an intriguing example reported by Marslen-Wilson, Levy, and Tyler (1982), narrators telling a story while pointing to comic-strip characters first looked away from their addressee, down towards their hands. Only then did they point. In other words, they were using their gaze to guide their addressees' gaze to their hands, and then using their hands to point to the referent of their speech.

Several studies have shown that knowing the addressee's current focus of attention is an important resource in conversation. In such situations, speaker and addressee share a joint focus of attention. They are both looking at the same region of visual space, and are aware that the other is also doing so. The number of potential referents is reduced to a subset of all possible referents (Beun & Cremers, 1998; Schmauks, 1991), which allows participants to use reduced verbal descriptions to identify an object. For example, in an experimental task where a director instructs a matcher as to how to build a model from blocks of several colors, an utterance such as "take a red block" will be ambiguous if there are several red blocks among the pieces to be used. But if the subset of blocks within the current joint focus of attention only includes one red block, then this reduced utterance is sufficient. In collaborative physical tasks in mediated settings, having access to information about a partner's gaze (for example by seeing a partner's eye movements superimposed on a shared document) results in more efficient communication (Kraut, Fussell, & Siegel, 2003; Velichkovsky, 1995). Gaze may also facilitate communication by allowing speakers and addressees to monitor each other and repair utterances as they are produced (Clark & Krych, 2004). Of course, analogous processes operate within language, as when focus spaces constrain pronominal reference (Brennan, 1995; Grosz & Sidner, 1986).

These accounts of the function of pointing gestures contrast with a "standard" view (Lyons, 1982) of deixis. According to this view, a pointing gesture serves to identify the referent of a deictic expression such as *that's my car*. Focusing attention does not play any significant role in this account. This assumption is embodied some experimental studies. For example, in a situation where a child has to pick out one present among four possible toys, a pointing gesture is insufficient to discriminate among them given that the toys are all placed too close to each other. In such situations, pointing has been described as less versatile, flexible and effective than language: "Linguistic devices, being the most versatile ones, seem to make other forms superfluous" (Pechmann & Deutsch, 1982, p. 331). Although this may be true for experimental settings where the task is to identify a target within a well-specified referential domain, a more typical function of pointing in naturalistic conversation (and one where it may be more effective) may be precisely to *construct* such a situation by focusing attention on a reduced domain of referents. This line of reasoning is of course derived from the joint-focus-of-attention view explained above.

How can these two accounts be reconciled? When is pointing used to focus attention, thereby indirectly facilitating reference, and when is it used to directly identify or locate a referent? To answer this question, we draw on recent proposals that view referring as a composite signal (Clark, 1996; Clark & Bangerter, 2004; Engle, 1998; see also Bavelas & Chovil, 2000), i.e., as typically consisting of both descriptive and indexical components. This proposal is consistent with observations from field research (Goodwin, 2003). Consider the following example (Schegloff, 1984, p. 280):

(1) Frank: why:nchu put that t the end uh the ta:ble there [pointing]

Here, Frank uses a combination of describing (*t the end uh the ta:ble*) and indicating (*there* accompanied by a pointing gesture) to get his addressee to recognize where to put a dish. What determines the combinations of describing and indicating that are used in referring? An important aspect of the composite signal view is that referring is flexible in both production and comprehension. In other words, speakers are capable of adapting their messages mid-utterance to fit addressees' evolving signals of comprehension (Clark & Krych, 2004). Likewise, addressees are capable of rapidly integrating speaker utterances with other sources of information (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002). The relative importance of describing and indicating in a referring act can be opportunistically adapted to the situation. In short, composite signals exhibit a tradeoff between describing (typically accomplished linguistically, except in the case of iconic gestures) and indicating (typically accomplished gesturally). Something similar holds for comprehension: People rely differently on linguistic and gestural components in comprehension depending on their relative ambiguity (Thompson & Massaro, 1986£). Thus, the informativeness of a pointing gesture may be limited by its ambiguity. There is an upper limit on the accuracy with which people can detect where another person is pointing (Bangerter & Oppenheimer, 2006). If the situation is ambiguous, pointing (indicating) alone will be insufficient and will need to be augmented by describing the target. People might then use a pointing gesture to direct their addressees' gaze to the target region and subsequently describe the object to pick it out among potential confounding referents. This may especially be the case when the referents of pointing gestures are distant from the pointer (Bangerter, 2004, van der Sluis & Krahmer, 2004). Therefore, we propose that the relative reliance on describing and indicating will vary as a function of the referring situation, and especially as a function of the ambiguity of pointing.

1.3 THE PRESENT STUDY

In the work reported here, we manipulated partner visibility and ambiguity to study their effects on the use of describing and indicating in a dyadic referential communication task. Our goal in manipulating visibility was to understand to what degree pointing gestures are used to communicate. The research summarized above indicates that in order to be considered communicative, gesture must be shown to be sensitive to context (e.g., visibility) in production as well as to have an effect on comprehension (Krauss et al., 1991). In other words, gestures should be produced more frequently and should result in more efficient verbal communication when partners are mutually visible than when they are not. Of particular interest are residual gestures produced in the non-visible condition. Many studies have found that gestures are not entirely suppressed when partners are not mutually visible (Alibali et al, 2001). Rather than using these findings to fuel the debate about whether gestures do or do not communicate, or which types of gestures do or do not communicate, it may be more profitable to analyze variations within types, across situations to see *when* gestures do or do not communicate (Bavelas, 1994). To this end, after an initial examination of the videotapes from our experiment, we decided to distinguish between those pointing gestures where the pointer's elbow was raised off the table and fully extended (hereafter: *full* points) and those where only the forearm was extended, the elbow remaining on the table (hereafter: *partial* points). Full points involve coordinated, purposeful movement of the arm from a resting position in the direction of the target. Thus, it seems likely that they are produced with communicative intent. Partial points do not involve as much movement and may be relatively automatic in production.

Our goal in manipulating ambiguity was to determine how it affects the relative use of describing and pointing in referential communication. When gestures are unambiguous, speakers should rely on them to a larger extent relative to describing. In an extreme case, a referent might be identified simply by pointing to it. When gestures are ambiguous, speakers should rely relatively more on describing a referent to identify it.

2 METHOD

2.1 PARTICIPANTS AND PROCEDURE

Twenty-four French-speaking pairs (director and matcher) worked together. They were seated side-by-side facing a large board supporting an array of portrait photos of people. The array was fully visible to both of them. A plexiglass panel was placed approximately 25 cm from the array to keep participants from touching targets. The task required the director to identify four target photos for the matcher. They talked and/or gestured freely to identify each target. Each target had a name that the director read to the matcher. The matcher wrote the name down on an answer sheet. After all four targets had been identified, the experimenter replaced the array with the next one. They identified targets from 12 arrays.

We manipulated the density of pictures in an array within subjects, and thus the ambiguity of gestural information. There were 2 sets of arrays with 8, 9, 11, 14, 20 and 37 pictures respectively. With four targets, these numbers constitute a 6-point linearly decreasing scale of the average probability of chance identification of a target. For example, the average probability of chance identification is 15.9% in the eight-picture array, 13.6% in the nine-picture array, 10.7% in the eleven-picture array, and so on. The number of pictures defines the density of the array. The denser the array, the closer pictures are to each other, and the more ambiguous gestures will be. Density is our operationalization of ambiguity. Pictures were arranged in a cloud-like fashion in the array, so as not to form obvious rows and columns. The order of presentation of arrays was counterbalanced.

We also manipulated visibility of partners to one another between subjects. In a visible condition (*n* = 12), pairs could see each other and thus use gestures; in a hidden condition (*n* = 12), they couldn't. The hidden condition was created by inserting a large wooden board between the director and matcher. The board completely hid them from each other, but did not obscure their view of the arrays.

Thus, the study had a 2 (visibility) by 6 (ambiguity) mixed-model design. The setup is shown in Figure 1.

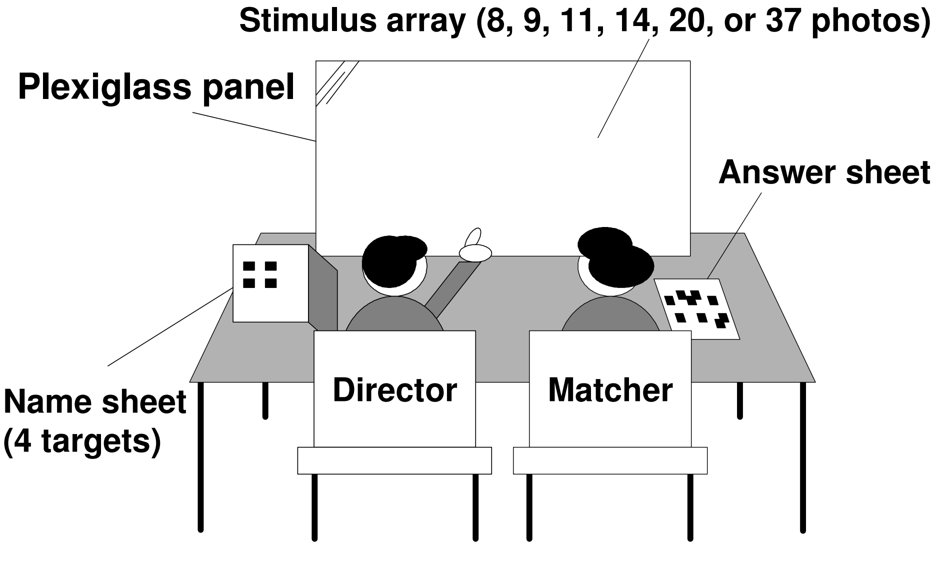


Figure 1: The set-up of the study.

2.2 DATA PREPARATION

Pairs were videotaped with two cameras. Videotapes were mixed onto a split-screen video. Communication was transcribed. Use of gestures and descriptions was coded. Descriptions included absolute spatial descriptions (e.g., *John is on the left*), spatial descriptions relative to another picture (*relative* spatial descriptions, e.g., *Betty is the one below the redhead*), feature descriptions (e.g., *John has glasses*) and deixis (e.g., *right here*, *over there*).

Pointing gestures were defined as a movement of the hand from a resting position (Sacks & Schegloff, 2002) in the direction of the array combined with partial or full extension of one or more fingers. In this respect, they were easy to distinguish from iconic gestures, where the hands were often raised towards the face of the speaker to mimic some feature being described (e.g., downward movement of both hands along the side of the head to mimic long hair). Typical resting positions of pointing gestures were the table or the face (some participants repeatedly touched their faces, especially their mouth or nose). Gesture performance is often decomposed into three phases: preparation, stroke and retraction (Kendon, 1972). The preparation phase involved raising the forearm from the table or removing the hand from the face and moving it in the direction of the array. Depending on the type of gesture, the elbow was sometimes raised off the table. The stroke phase typically involved holding the point for some amount of time (e.g., until an acknowledgment from the matcher) or hand waving, finger extension, flicking or another kind of gesticulation. The retraction phase involved returning the arm to the previous resting position or to another one.

We coded two kinds of pointing gesture, partial and full. Partial pointing gestures were defined as an extension of the arm in which the elbow remained on the table. Full pointing gestures were coded when the elbow was raised off the table. There was a degree of variation in the amount of extension of the upper arm and forearm, typically depending on how much distance participants wanted to cover. Participants sometimes augmented full pointing gestures with fully extended arms by leaning forward in their chairs, especially when pointing at peripheral targets. Examples of partial and full pointing gestures are shown in Figure 2. Both partial and full pointing gestures often involved hand movements at the moment of the stroke.

Inter-rater agreement was assessed by having two independent coders double-code the number of times each type of verbal description and gesture was used per target (irrespective of who used it). Four pairs were double-coded for verbal descriptions and three for gestures. Correlations were computed. They varied between .87 and .93 for verbal descriptions with (all *p*s < .0001), indicating excellent agreement. Correlations were .82 for partial points and .92 for full points (both *p*s < .0001).

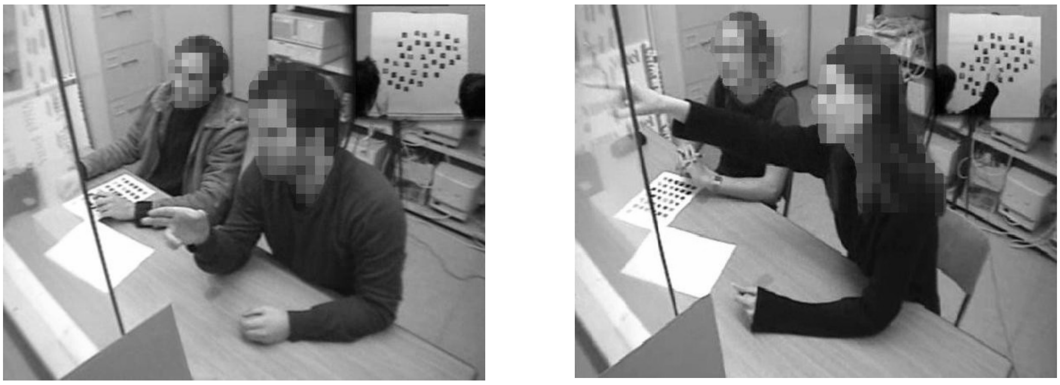


Figure 2: Examples of partial (left side of figure) and full (right side of figure) pointing gestures. The insert in the top right-hand corner of each image depicts an over-the-shoulder view of the task situation.

3 RESULTS

We first present descriptive analyses and examples (3.1). We then analyze the relative use of full and partial points (3.2), verbal effort (3.3) and verbal descriptions (3.4).

3.1 DESCRIPTIVE ANALYSES AND EXAMPLES

Pairs varied in their use of gestures, even within conditions. Some visible pairs used gestures for almost every trial, whereas others did not. Most gestures and descriptions were produced by directors. Identifying targets clearly was easier in the visible condition than in the hidden one. Here is an example of how descriptions were used in the visible condition. Participants are identifying Rachel, on one of the 37-picture arrays (i.e., the highest ambiguity level). Descriptions coded are indicated within parentheses, with the abbreviated type of description in subscript (RLOC = relative location description, ALOC = absolute location description, FEAT = feature description).

1 D Rachel um with (the one next um)$_{\text{RLOC}}$
2 M (next to the blonde?)$_{\text{RLOC}}$
3 D yeah that's it (with the blue background)$_{\text{RLOC}}$

In this example, Rachel is identified with three relative location descriptions and a pointing gesture. The pairs use a presumably more salient adjacent picture, a blonde (2) with a blue background (3), in order to get to her. Below is a pair in the hidden condition trying to identify the same picture, also using a relative location description to begin with.

1 D (she's on your side)$_{\text{ALOC}}$ you see (there's a girl with a blue background with her head raised who's laughing the head a bit back ha ha ha)$_{\text{RLOC}}$ (1.5) you see?
2 M oh yeah yeah
3 D well (just on the side)$_{\text{RLOC}}$ (she's she has her face on the side)$_{\text{FEAT}}$ (with a background a bit of red there)$_{\text{FEAT}}$
4 M yeah
5 D (and then a smile in the the corner)$_{\text{FEAT}}$

Although this pair uses the same strategy, the director and matcher expend more effort in grounding each step of the identification (Clark & Krych, 2004). The director first uses an absolute location description (*she's on your side*) to focus attention on the target region. He then describes the adjacent picture in much more detail (1) before mentioning the target (3, 5). The director also makes sure the matcher is looking at the right picture (*you see*?) before proceeding. In the visible example, there is no feature description, whereas in the hidden example, there are three. In other words, the hidden pair also invests more effort in verifying that the target is the correct one after having tentatively identified it. This is similar to the overspecification observed by van der Sluis and Krahmer (2004) in a production task.

3.2 RATES OF USE OF FULL AND PARTIAL POINTS

We computed the rate of gestures per 100 words (used by both director and matcher) as a dependent measure (Alibali et al., 2001). A higher rate indicates that gestures are used more relatively to words. A lower rate indicates relatively more reliance on words (see Table 1). Visible pairs used full points at a higher rate than hidden pairs. They used less full points per 100 words as ambiguity increased. A 2 (visibility) by 6 (ambiguity) mixed-model ANOVA revealed a main effect of visibility, $F(1,165) = 132.71$, $p < .0001$, a main effect of ambiguity, $F(5,161) = 10.07$, $p < .0001$, and an interaction, $F(5,161) = 10.82$, $p < .0001$. The main effect of ambiguity does not show whether there is an overall increase or decrease, but we tested for a linear trend for visible pairs. It was significantly negative, $F(1,1112) = 15.47$, $p = .0001$. This suggests that full pointing gestures are indeed produced with communicative intent, because visible pairs relied on them more than hidden pairs. It also suggests that pairs relied relatively more on verbal information and less on gestures as ambiguity increased. The rate of partial points per 100 words varied according to ambiguity, but it was not significantly different for visible and hidden pairs: A 2 (visibility) by 6 (ambiguity) mixed-model ANOVA revealed only a main effect of ambiguity, *F*(5,165) = 3.28, *p* = .008. This suggests that partial points are not used to communicate, as their use is not sensitive to visibility.

	Ambiguity (Number of pictures in array)					
	8	9	11	14	20	37
Visible condition						
Full points	7.43 (7.59)	6.34 (6.49)	8.15 (11.9)	5.94 (4.19)	5.32 (3.63)	3.61 (2.41)
Partial points	0.45 (1.80)	1.27 (3.20)	0.55 (2.01)	0.44 (1.96)	0.21 (0.88)	0.40 (1.27)
Absolute location	2.79 (5.17)	5.14 (5.04)	4.47 (5.24)	2.53 (3.16)	2.33 (3.24)	1.73 (2.05)
Relative location	1.51 (4.06)	1.76 (5.05)	1.57 (3.31)	1.73 (4.16)	1.90 (3.77)	2.35 (3.20)
Feature	7.43 (5.29)	5.69 (5.26)	5.47 (4.87)	8.58 (3.73)	7.35 (3.95)	7.16 (4.06)
Deixis	3.88 (5.35)	3.43 (4.59)	3.76 (4.59)	3.87 (3.83)	3.78 (3.83)	2.78 (3.02)
Hidden condition						
Full points	0.04 (0.32)	0.17 (0.79)	0.31 (0.95)	0.00 (0.0)	0.10 (0.53)	0.13 (0.46)
Partial points	0.24 (0.98)	0.54 (1.60)	0.27 (1.02)	0.44 (1.14)	0.26 (0.74)	0.42 (0.95)
Absolute location	3.82 (3.07)	4.24 (3.97)	3.09 (3.27)	2.90 (2.59)	2.51 (2.24)	2.68 (2.10)
Relative location	1.82 (2.82)	2.37 (4.15)	2.10 (2.61)	2.23 (2.59)	1.48 (1.96)	2.43 (2.26)
Feature	7.14 (3.53)	7.28 (4.65)	6.83 (3.56)	7.79 (3.19)	7.48 (2.85)	5.74 (2.52)
Deixis	0.02 (0.17)	0.09 (0.77)	0.00 (0.0)	0.03 (0.23)	0.03 (0.22)	0.03 (0.27)

Table 1: Mean rates per 100 words of full and partial points, absolute and relative location descriptions, feature descriptions and deixis as a function of visibility and ambiguity.

3.3 VERBAL EFFORT

Mutual visibility reduced verbal effort, and ambiguity increased it. A 2 (visibility) by 6 (ambiguity) mixed-model ANOVA with mean number of words per target as dependent variable revealed a main effect of visibility, *F*(1,165) = 141.75, *p* < .0001, a main effect of ambiguity, *F*(5,161) = 38.75, *p* < .0001, and an interaction, *F*(5,161) = 6.44, *p* < .0001. Figure 3 shows the mean levels and error bars, as well as significant differences (identified by repeated-measures contrasts) between consecutive data points of the same line. The fact that visible pairs needed fewer words to complete the task suggests that pointing gestures do indeed reduce verbal effort. However, the reduction in verbal effort could also be due to confounding factors, such as the fact that visible pairs were much more aware of where their partners were looking. Another, more direct, way of testing whether gesture use aids referential communication is by looking at differences between pairs in the same condition. We correlated the total number of gestures of both types with the total number of words for each pair (*n* = 12). Visible pairs that used more full points used less words, *r* = -.81, *p* < .001, but their total number of partial points was unrelated to the total number of words, *r* = .33, *ns*. Thus, the lower verbal effort in the visible condition is at least partly related to gesture use. For hidden pairs (*n* = 12), we found that their total number of full points was not related to their total number of words, *r* = -.16, *ns*. However, their total number of partial points was *positively* related to their total number of words, albeit marginally, *r* = .57, *p* = .052.

3.4 VERBAL DESCRIPTIONS

Because verbal effort varied according to ambiguity and visibility, we computed the mean number of times verbal descriptions (absolute location descriptions, relative location descriptions, feature descriptions) were used per 100 words (irrespective of who used them). This is their rate of use or *relative* use. We also computed the rate of use of deictic expressions (e.g., *here*, *there*). Descriptive data are shown in Table 1. Feature descriptions were most often used, followed by absolute location descriptions and relative location descriptions. Deixis was used regularly only by visible pairs, suggesting that they accompanied pointing gestures. In what follows, we report inferential statistics for each type of verbal description.

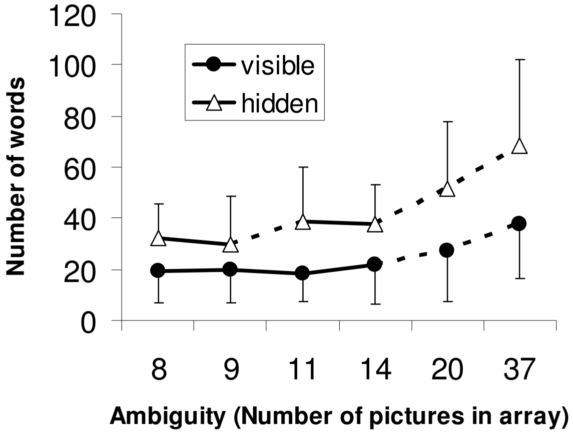


Figure 3: Mean number of words per target as a function of visibility and ambiguity. Error bars indicate one standard deviation. Dotted lines indicate significant differences between consecutive points.

The relative use of absolute location descriptions varied according to ambiguity. A 2 (visibility) by 6 (ambiguity) mixed-model ANOVA revealed a main effect of ambiguity, *F*(5,161) = 10.38, *p* < .0001, and an interaction between ambiguity and visibility, *F*(5,161) = 3.11, *p* = .01. The main effect of visibility did not reach significance, *F*(1,165) = .03, *ns*. We tested for linear and quadratic trends in the relationship between absolute descriptions and ambiguity. There was a significant quadratic trend for visible pairs, *F*(2,568) = 14.58, *p* < .0001, adjusted *R2* = .045, i.e., absolute location descriptions were used relatively less often at low and high levels of ambiguity than at intermediate levels. For hidden pairs, there was a significantly decreasing linear trend, *F*(1,541) = 25.9, *p* < .0001, adjusted *R2* = .044.

The relative use of relative location descriptions did not vary significantly according to density or ambiguity, all *F*s < 1.6, *ns*.

The relative use of feature descriptions varied according to ambiguity. A 2 (visibility) by 6 (ambiguity) mixed-model ANOVA revealed a main effect of ambiguity, *F*(5,161) = 7.50, *p* < .0001, and an interaction between ambiguity and visibility, *F*(5,161) = 4.70, *p* = .0001. The main effect of visibility did not reach significance, *F*(1,165) = 0.12, *ns*. There was a significant quadratic trend for hidden pairs, *F*(2,540) = 4.28, *p* = .014, i.e., feature descriptions were used relatively less often at low and high levels of ambiguity than at intermediate levels. However, the size of the effect was small, adjusted *R2* = .012.

The relative use of deictic expressions varied only according to visibility. A 2 (visibility) by 6 (ambiguity) mixed-model ANOVA revealed a main effect of visibility, *F*(1,165) = 164.23, *p* < .0001, with more deictic expressions being used by visible pairs. There was no effect of ambiguity, *F*(5,161) = 1.40, *ns*, nor was there a significant interaction, *F*(5,161) = 1.45, *ns*.

Thus, taken together, the rate of relative use of verbal descriptions only varied substantially for absolute location descriptions at different levels of ambiguity. Both hidden and visible pairs used them less often at high levels of ambiguity. Visible pairs increased their use from lower to intermediate levels of ambiguity.

4 Conclusions

The most important finding relates to the distinction between two types of pointing gestures, partial and full points. Pairs used more full points when their partners could see them than when they couldn't. This suggests that full points are produced with communicative intent. Pairs that used more full points also used fewer words to complete the experiment, indicating that full points affect comprehension. Finally, the relative reliance on full points decreased with increasing ambiguity. This is consistent with the composite signal view of multimodal communication: people rely differentially on linguistic and gestural components of a multimodal signal in an opportunistic manner.

The use of partial points was not sensitive to visibility, suggesting that their production is not intended to communicate and may be automatic. As was discussed above for other types of gesture, partial points may possibly be functional for the speaker. They may serve as a visual marker. Some directors looked back and forth between their list of targets and the target on the board, possibly to verify whether they had the correct picture before describing it to the matcher. A partial point may have helped them remember where the target was located on the array, especially for dense arrays. Alternatively, partial points may also reflect difficulties in production. The fact that they were marginally positively correlated with the number of words used by hidden pairs is consistent with this interpretation. Hidden speakers that were having difficulty formulating a description may have spontaneously produced a partial point, much in the same way as they spontaneously produced iconic gestures during describing (Morrel-Samuels & Krauss, 1992). Visible speakers in the same situation may have simply extended their arm to transform the partial point into a full point.

The relative use (i.e., per 100 words) of different verbal descriptions did not vary much with ambiguity. This means that pairs did not change their verbal strategies fundamentally as ambiguity increased. Instead, they simply did more of the same thing. The only reliable trend was a curvilinear relationship between ambiguity and use of absolute location descriptions (e.g., descriptions like *in the middle* or *on your side*). At intermediate levels of ambiguity, absolute location descriptions were used more often than at low and at high levels of ambiguity. Although this finding is difficult to interpret, it may be that absolute location descriptions are unnecessary at lower levels of ambiguity because a pointing gesture suffices to focus attention on the approximate region of the target. This is consistent with the finding that only visible pairs exhibited the curvilinear relationship: only visible pairs are in a position to substitute pointing gestures for absolute location descriptions and thus show a lower rate of use at lower levels of ambiguity. At higher levels of ambiguity, absolute location descriptions may be less useful because more pictures may be in the same approximate target region. This limitation on the efficiency of absolute location descriptions at higher ambiguity levels holds for both visible and hidden pairs and is therefore consistent with the observed data.

The data support the idea that gestures within a given type (here, pointing gestures) may have both communicative and speaker-related functions. More generally, the data show that it may be worth exploring how the precise function of a gesture varies according to the communicative context (Alibali et al., 2001, Bavelas, 1994). It is entirely possible that a full point may start out as a partial point. Moreover, we observed (as did van der Sluis & Krahmer, 2004) that many pointing gestures exhibited some kind of lateral hand movement at the stroke. Informal observation suggested that these movements may be similar to iconic gestures in that they may be related to the production of verbal descriptions. A single pointing gesture may therefore have several different functions at different moments in its production. Taxonomic approaches to gesture may indeed be potentially misleading, as suggested by Bavelas (1994). To conclude, we suggest that more research should attend to the possible variations in gesture function within types and in different communicative contexts.

REFERENCES

Alibali, M.W., & Di Russo, A.A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development, 14*, 37-56.

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production. Some gestures are meant to be seen. *Journal of Memory and Language, 44*, 169-188.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15*, 415-419.

Bangerter, A. & Oppenheimer, D. M. (2006). Accuracy in detecting referents of pointing gestures unaccompanied by language. *Gesture*, 85-102.

Bavelas, J. B. (1994). Gestures as part of speech: Methodological implications. *Research on Language and Social Interaction, 27*, 201- 221.

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning : An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology, 19*, 163-194.

Beattie, G. W., & Shovelton, H. K. (2005). Why the spontaneous images created by the hands during talk can help make TV advertisements more effective. *British Journal of Psychology, 96*, 21-37.

Beun, R.-J., & Cremers, A. H. M. (1998). Object reference in a shared domain of conversation. *Pragmatics & Cognition, 6*, 121-152.

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes, 10*, 137-167.

Buchler, J. (Ed.) (1940). *Philosophical writings of Peirce*. London: Routledge and Kegan Paul.

Bühler, K. (1965). *Sprachtheorie*. [Theory of language.] Stuttgart: G. Fischer.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., and Carlson, G. N. (2002). Circumscribing referential domains in real-time language comprehension. *Journal of Memory and Language, 47*, 30-49.

Chawla, P., & Krauss, R. M. (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology , 30*, 580-601.

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

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: where language, culture and cognition meet* (pp. 243-268). Hillsdale, NJ: Lawrence Erlbaum.

Clark, H. H., & Bangerter, A. (2004). Changing ideas about reference. In I. Noveck & D. Sperber (Eds.), *Experimental Pragmatics* (pp.25-49). Basingstoke: Palgrave Macmillan.

Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speech. *Cognition* , *84*, 73-111.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language , 50*, 62-81.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J. F. LeNy & W. Kintsch (Eds.), *Language and comprehension* (pp. 287-299). Amsterdam: North-Holland.

Engle, R. A. (1998). Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*.

Goodwin, C. (2003). Pointing as situated practice. In S. Kita (Ed.), *Pointing: where language, culture and cognition meet* (pp. 217-241). Hillsdale, NJ: Lawrence Erlbaum.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology, 10*, 57-67.

Grosz, B., & Sidner, C. (1986). Attentions, intentions and the structure of discourse. *Computational Linguistics, 12*, 175-204.

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

Kendon, A. (1972). Some relationships between body motion and speech. An analysis of an example. In A. Siegman & B. Pope, (Eds.), *Studies in Dyadic Communication* (pp. 177-210). Elmsford, New York: Pergamon Press.

Kendon, A. (1994). Do Gestures Communicate? A Review. *Research on Language and Social Interaction 27*, 175-200.

Kita, S. (Ed.) (2003). *Pointing: where language, culture and cognition meet*. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures*. Journal of Experimental Social Psychology , 31*, 533-552.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction, 18*, 13-49.

Louwerse, M. M., & Bangerter, A. (2005). Focusing attention with deictic gestures and linguistic expressions. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society.*

Lyons, J. (1981). *Language, meaning and context*. Fontana Paperbacks.

Marslen-Wilson, W., Levy, E., & Tyler, L. K. (1982). Producing interpretable discourse: The establishment and maintenance of reference. In R. J. Jarvella & W. Klein (Eds.), *Speech, place and action. Studies in deixis and related topics*. (pp. 339-378). Chichester: John Wiley.

Melinger, A., & Levelt, W. (2004). Gesture and the communicative intention of the speaker. *Gesture, 4*, 119-141.

Morrel-Samuels, P., & Krauss, R.M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 615-623.

Morsella, E., & Krauss, R. M. (2004). The role of gestures in spatial working memory and speech. *American Journal of Psychology, 117*, 411-424.

Özyürek, A. (2002). Do speakers design their co-speech gestures for their addresees? The effects of addressee location on representational gestures, *Journal of Memory and Language, 46*, 688-704.

Pechmann, T., & Deutsch, W. (1982). The development of verbal and nonverbal devices for reference. *Journal of Experimental Child Psychology, 34*, 330-341.

Rimé, B. (1982). The elimination of visible behavior from social interactions: effects on verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology, 12*, 113-129.

Rimé, B., & Schiaratura, L. (1991). Gesture and speech. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of Nonverbal Behavior* (pp. 239-281). Cambridge: Cambridge University Press.

Schegloff, E.A. (1984). On some gestures' relation to talk. In J.M. Atkinson & E. J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 266-296). Cambridge: Cambridge University Press.

Schmauks, D. (1991). *Deixis in der Mensch-Maschine-Interaktion*. [Deixis in human-computer interaction.] Tübingen: Max Niemeyer.

Thompson, L. A., & Massaro, D. W. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology, 42*, 144-168.

Van der Sluis, I., & Krahmer, E. (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. *Proceedings of the ICSLP-2004*.

Velichkovsky, B. M. (1995). Communicating attention. Gaze position transfer in cooperative problem solving. *Pragmatics and Cognition, 3*, 199-224.

Generating Text, Diagrams and Layout Appropriately According to Genre

John A. Bateman and Renate Henschel
Faculty of Linguistics and Literary Sciences
University of Bremen
Bremen, Germany
{bateman|rhenschel}@uni-bremen.de

Abstract

In this paper, we summarize the consequences for an integrated generation of text, diagram and layout that we have drawn from our ongoing empirical studies of the constraints imposed by multimodal genres. Although techniques are now available for producing diverse multimodal representations, this very flexibility can become a liability unless it is controlled appropriately for the purposes a representation is attempting to achieve. We see a precise definition of multimodal genre as the most general way of achieving such control, set out a prototypical implementation, and discuss some currently important theoretical and technical issues for further development.

Keywords: Page layout, document design, multimodal document generation, natural language generation.

1 Introduction

In this paper, we summarize the consequences for the integrated generation of text, diagram and layout that we have drawn from our ongoing empirical studies of the constraints imposed by document types or, as we prefer to term them, multimodal genres. For the genres that we are examining and trying to include within the generation capabilities of our prototypes, we typically find that substantial information is expressed spatially in the visual array rather than in details of macro-punctuation within flowing text.

In Bateman et al. (2001), a process was set out by which multimodal rhetorical structures as originally proposed in work such as André et al. (1993) and de Carolis et al. (1997), could be used to drive a variety of page presentations varying widely in layout. This raised explicitly the problem of how to constrain the process so that documents 'appropriate' for specific purposes could be produced. Techniques are now becoming available for producing diverse multimodal representations but that very flexibility will itself become a liability unless it is controlled appropriately for the purposes that a representation wants to achieve. Formulating an account in which this issue could be addressed empirically then became the main goal of our ESRC project 'Genre and Multimodality' (cf. Allen et al., 1999; Bateman et al., 2004 and `http://purl.org/net/gem`).

One principal result of this work was the proposal that a precise definition of multimodal genre will allow us to achieve appropriate control of the multimodal generation process. Multimodal genre provides a space of possibilities in which multimodal documents can be positioned according to the realization options they take up—analogously to how registers of written and spoken language can be distinguished. We accordingly characterize multimodal genres in terms of collections of features specified at several distinct layers of description. Generating appropriate page layouts then becomes a process of relating the rhetorical organisation and the layout structure of a document while at the same time enforcing configurations of multi-layer combinations of features that have been found to be appropriate for particular document genres.

The empirical basis of this approach was a collection of multimodal documents selected from several distinct multimodal genres or text types. This served as a source of systematic and motivated constraints for sophisticated layout generation. We are now continuing this work collecting further examples of multimodal documents in order both to extend our empirical basis and to test our proposed mechanisms for document generation. As an example of the latter, we also present here a prototype implementation of a general algorithm for transforming a multimodal rhetorical structure into a not necessarily isomorphic layout structure dependent on features of the chosen genre. We illustrate the prototype with two different genres drawn from the GeM corpus, bird guides and instruction manuals, and discuss some of the technical issues that arise more generally.

2 Modelling Layout

Descriptions of document layout can be usefully classified according to whether they start from a notion of *text* or from a notion of the *page*. The former sees layout as more or less extreme diversions from the inherent linearity of text; the latter sees layout as a visual phenomenon that is inherently two dimensional (on the page) and spatial. In natural language generation we find representatives of both directions.

Some of the earliest work in the first direction includes Hovy and Arens's (1991) addition of LaTeX commands, such as enumeration, bullet lists, and emphasis, to their generator's output and Sefton's (1990) extension of a generation grammar with a graphological level including punctuation and some low-level formatting similar to that of Hovy and Arens. Many systems now employ similar methods to produce text that is more visually informative than unadorned sequences of characters. Here the widespread use of web-browsers capable of turning HTML files into laid-out pages has played an important role in promoting HTML as an output format for NLG systems (cf. Kruijff et al., 2000). The spatial approach, in contrast, is represented by the tradition begun by André et al. (1993), in which a multimodal presentation plan is derived and this is then rendered by more or less complex page layout algorithms which try and respect the logical neighbourhood relations of the presentation plan in terms of spatial neighbourhoods (e.g., Graf, 1995; Feiner, 1988).

In both traditions it has become clear that it is not in general sufficient for layout, be it textual or visual, to simply take over the rhetorical organisation of the presentation plan. In order to investigate this phenomenon more effectively and to build appropriate degrees of controllable flexibility into our generation systems, it is necessary to characterise at least two kinds of document description—one oriented to the layout and one to the rhetorical purpose—independently of one another. Only by this means is it possible to explore the limits of the mappings between them.

To support this investigation, we formulated in the course of our corpus work a multi-layer annotation scheme for multimodal documents that are without animation and which use a page-metaphor presentational style. The layers of this model are described in more detail in, for example, Delin et al. (2002/3) and Bateman et al. (2004); here we concentrate on just two, the presentational layer for layout itself and the rhetorical layer. The rhetorical layer is a multimodally extended version of Mann and Thompson's (1988) rhetorical structure theory (RST). The layout layer is defined as follows.

2.1 The GeM Layout Units

In typography, the minimal layout element (in text) is the glyph. In the GeM project, we were primarily concerned with typographical and formatting effects at a more global level, and so consider as minimal layout elements text blocks of the paragraph level, pictures in their entirety, and all other layout elements which are differentiated as a whole from their environment *visually*. We call these minimal layout elements **layout units**; the current catalogue of these units is shown in Table 1.

Typographical realization.

The most obvious difference to be observed in realized layout units is the mode in which they are realized—typically linguistic or graphical. Dependent on the chosen mode, different sets of

continuous homogeneous text block	spatially distinguished sentence fragments initiating a list	headings, titles, headlines
photos, drawings, diagrams, figures (without caption)	captions of photos, drawings, diagrams, tables	text in photos, drawings, diagrams
icons	table cells	list headers
list items	list labels (itemizers)	items in a menu
page numbers	footnotes (without footnote label)	footnote labels
running heads	emphasized text, i.e., text which in some way stands out by size, type face, or weight from its background	horizontal or vertical lines which function as delimiters between columns or rows
lines, arrows, polylines which connect other layout units		

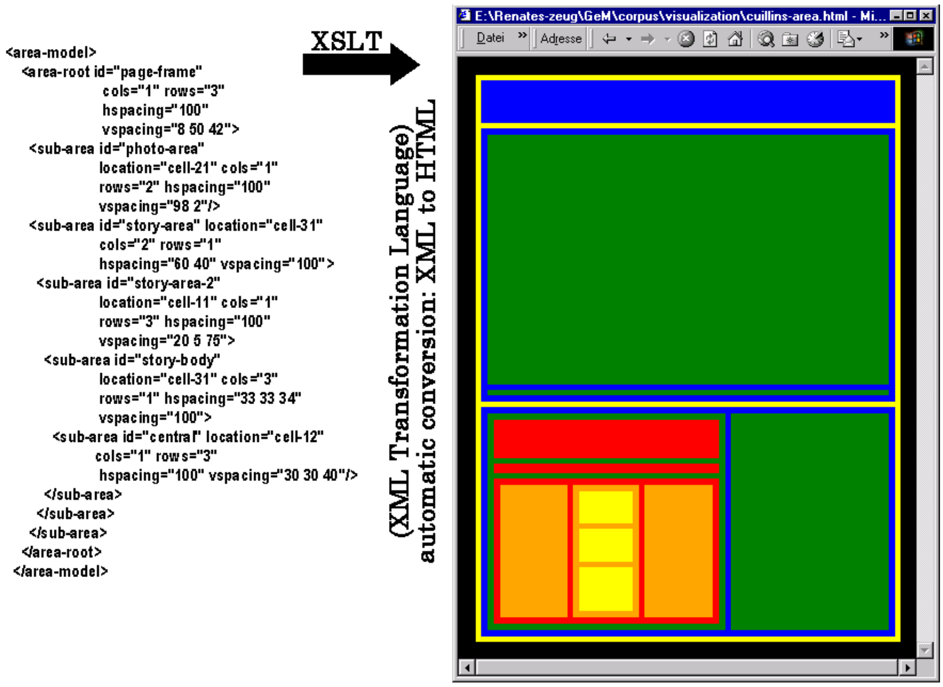
Table 1: Layout units of the GeM model. Each of these identifiable elements when found on a page to be analysed is considered as a layout unit and is assigned a set of properties distinctive to its type.

features describe other layout characteristics. For textual elements, we consider font family, font size, font weight, font style (italics or not), justification, color, case, and so on. For the graphical layout units, the only choice we have for their realization presently is their size, because we are currently working with ready-to-show pictures as input; straightforward extensions would include standard classifications of graphical elements. All such features are included as attributes of appropriate elements in the layout structure.

Layout structure.

The layout structure describes how layout units are hierarchically grouped into larger layout chunks. For instance, the heading and its associated text form together a larger layout element, or the cells of a table form the larger layout element “table”. The criterion for grouping layout units into chunks is that the chunk should consist of elements of the same visual realization (font-family, font-size, . . .), or the chunk is differentiated as a whole from its environment *visually* (e.g. by background colour or a surrounding box). Some motivations and methods for identifying layout chunks have been discussed by, for example, Reichenberger et al. (1995), Summers (1998), Eglin and Bres (2004) and others; we are also currently planning to extend these methods using results from eyetracking experiments such as those reported in Holsanova et al. (2006). Any layout chunk can consist of layout elements involving different modes (text and graphics).

The layout structure of a document is a hierarchical structure, with the entire document being the root. Each layout chunk is a node in the tree, and the minimal layout units are the terminal nodes of that tree. The grouping into complex chunks—the layout structure—is determined by: (i) the rhetorical structure of the information to be presented; and (ii) **canvas constraints**—constraints arising from the medium used (paper size and quality) and from presentation decisions imposed on a document as a whole (cf Bateman et al., 2004). Examples for layout chunks derived from the RST structure are chapters, sections, and paragraphs. Examples for layout chunks generated by canvas constraints are pages and columns. Typical for this second kind of layout grouping is that even sentences are broken apart, and can readily belong to different layout chunks in the output document. The final layout which the reader sees then shows a layout structure subject to both types of constraints—i.e., RST constraints and canvas constraints.



```
<area-model>
  <area-root id="page-frame"
                cols="1" rows="3"
                hspacing="100"
                vspacing="8 50 42">
   <sub-area id="photo-area"
               location="cell-21" cols="1"
               rows="2" hspacing="100"
               vspacing="98 2"/>
   <sub-area id="story-area" location="cell-31"
               cols="2" rows="1"
               hspacing="60 40" vspacing="100">
    <sub-area id="story-area-2"
               location="cell-11" cols="1"
               rows="3" hspacing="100"
               vspacing="20 5 75">
     <sub-area id="story-body"
               location="cell-31" cols="3"
               rows="1" hspacing="33 33 34"
               vspacing="100">
      <sub-area id="central" location="cell-12"
              cols="1" rows="3"
               hspacing="100" vspacing="30 30 40"/>
     </sub-area>
    </sub-area>
   </sub-area>
  </area-root>
 </area-model>
```

XSLT
(XML Transformation Language)
automatic conversion: XML to HTML
E:\Renates-zeug\GeM\corpus\visualization\cuillins-area.html - Mi...
Datei
Adresse

Figure 1: GeM Area model and an automatic XSLT-based visualisation. This visualisation highlights the origin of the areas on the page by allocating a random colour to the area rather than its associated contents.

2.2 Page positioning: the GeM area model

The layout of a document is not fully determined by grouping layout units into a tree structure; further information is required about the actual position of each unit in the document (on or within its page). For this, we introduce an **area model**,[1] which recursively specifies rectangular sub-areas of the page area in a grid-like manner. These sub-areas then serve to determine the position of each layout chunk or layout unit. Two layout elements are called **adjacent**, if they are placed into two either horizontally or vertically adjacent subareas.

An example of both the XML source for an area model and an automatically produced visualisation of that model is shown in Figure 1. This is the area model of a newspaper page example that we discuss in some detail in Bateman et al. (2004). The XML characterization of the multi-level description as a whole has been given in Bateman et al. (2002), where we provide examples of the kind of annotation applied for all layers of the model and also contrast the approach with some other current multilayered XML-based approaches to corpus design. Probably the most crucial aspect of the design is its separation of annotation into distinct layers of standoff annotation in the manner proposed by Thompson and McKelvie (1997) and the reliance on a non-temporally organised layer of base units. This distinguishes the approach from some accounts of document description, where distinct kinds of information are by no means always so cleanly separated (e.g., Anjewierden, 2003).

2.3 Contrasting layout structure with document structure

In order to bring out the particular contribution of the layout structure, it is useful to contrast it more explicitly with the level of *document structure* proposed by Power *et al.* (Power et al., 2003; Power, 2000). Power *et al.*'s approach is, as mentioned above, situated within the textual

[1] This should not be confused with the similarly named but different construct from XSL-FO, although there are interesting similarities with proposals within the page media modules of CSS3 currently under development: `http://www.w3.org/TR/css3-page/`.

perspective on layout. They investigate particularly how distinct decisions concerning indentation, enumerated and itemized lists, groupings into sentences, clauses related by punctuation, etc. can interact with the linguistic phrasing required.

Power *et al.* use document structure to develop a flexible approach for transforming rhetorical structure into distinctly formatted possibilities on the page. Formally, document structure provides a hierarchical organisation that combines the kind of lower level information found in paragraphs and below—e.g., paragraphs as such, itemized lists, indented elements, punctuation proper—with the kind of larger scale 'logical structure' promoted in traditional markup languages such as SGML (cf. Goldfarb, 1990; Summers, 1995). At the highest levels in the hierarchy we therefore find a document decomposed into elements such as section, heading, body, etc. These are in turn decomposed further until we reach paragraphs, which are then themselves decomposed into elements derived from Nunberg's (1990) formal structural view of punctuation. One strong criterion for deciding whether an element is to be captured in the document structure is then whether or not it can influence the linguistic expression that is necessary.

The difference between this kind of structuring and that described above from the GeM perspective has already been characterised well by Bouayad-Agha (2001, p46). Whereas document structure focuses on a view of semantic content, pre-organised according to the aims of a particular document into sections and their subelements, the layout structure being described here is oriented more towards a visual orientation to documents that builds on the visual perceptual system involving Gestalt mechanisms for perceiving spatial configurations. For the task that Power *et al.* were setting themselves—particularly to uncover the dependencies between linguistic phrasing, connective choice, etc. and decisions of document layout—their document structure is clearly a sensible level of abstraction to work with; we will discuss below to what extent it might be useful for the GeM model to include similar information.

The aims originally pursued in the GeM project and now being taken further in our current work have been quite different. In particular, we were focusing from the outset on methods for moving beyond the largely linear views of layout that naturally dominate when layout is seen as an issue of text formatting and 'macro-punctuation'. Although this is common when starting from a linguistic perspective, there are actually many documents where this is not the organisation employed; we discuss this further in Section 5 below. The visual starting point for layout structure also requires the inclusion of a range of elements that are naturally excluded from document structure. In particular, layout structure must:

- reflect the production and canvas constraints which the realization of a given document structure is subjected to (decisions about pagination, columns, margins, hyphenation, etc.);
- specify access and navigational elements—layout elements which are not derived from the content, but which serve to guide the reader through the document (e.g. page numbers, pointers, running heads, titles);
- specify the position of layout elements on the page.

All of these considerations are mobilised in the service of constructing recognisable multimodal genres and so all need to find an explicit place in the account.

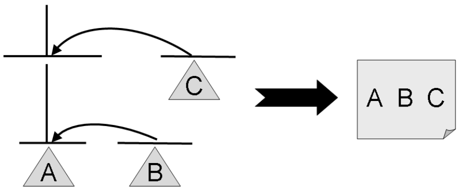
3 The Relation between RST Structure and Layout Structure

In its original form, RST investigates the relations which hold between the contents of consecutive clauses, or of bigger adjacent fragments of a text—the so-called text **spans**. RST structures also often function as data structures mediating between text planning and tactical generation in pipeline organized NLG systems; the terminal nodes of the RST tree are then semantic propositions. We adopt this latter model for the generation process developed here, generalizing it to hold over multimodal presentations in the manner proposed by André et al. (1993) and others mentioned above.

Now, however, as motivated empirically by Bateman et al. (2001), Bouayad-Agha et al. (2000) and others, it appears to be the case that a representation of a document that reflects visual grouping, regardless of whether this is 'logically' or 'visually' motivated, cannot be derived simply by maintaining the structure inherent in the rhetorical structure. The output layout structure does not in general preserve the RST input structure. To account for this, we see the process of shaping content for multimodal presentation as a generalization of linguistic linearization: here we are not, however, linearizing but 'spatializing' the input by *cutting* it at various places and allowing the created segments to flow into layout elements. Starting then from a multimodal rhetorical structure consisting of rhetorical relations between rhetorical units that may already have had a mode determination made (e.g., graphic, created text or semantic specification for text), we successively construct a corresponding layout structure by applying the following three principal types of structural transformation.

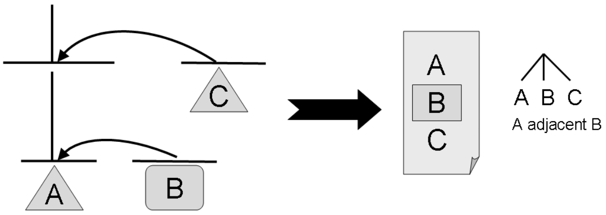
Sequential layout (concatenation).

The terminal nodes of an RST tree are all realized inside one and the same layout block (in case of text, with identical typography) maintaining the adjacency of nucleus/nuclei and satellites of one and the same relation. The rhetorical structure is not expressed typographically but may result in linguistic marking such as connectives.



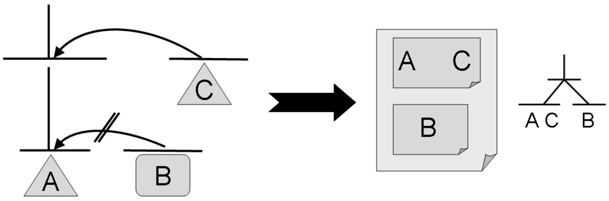
Emphasis.

A certain satellite or nucleus is realized with different layout properties than its sister nodes, thus creating an extra layout block, but maintaining adjacency with the other relation constituents (nucleus and satellites). Structure is preserved and the distinction is expressed only with layout properties. Thus 'emphasis' here should not be seen as a mechanism to highlight something against a context, but rather to distinguish constituents from one another. The 'emphasised' element stands out against the background in some way without being displaced from that background.



Extraposition.

A certain satellite or nucleus is cut from the RST tree and realized at a different place in the document, not necessarily adjacent to its sister nodes. Here, therefore, the structure is changed: layout structure does not reflect rhetorical structure. In this case, we often find pointers of various kinds to render the lost rhetorical relationships recoverable.



For multinuclear relations, we state that they should be of equal layout status: if one nucleus is cut, all are cut; if one nucleus is emphasized, all of them are emphasized.

We selected the name of this operation due to the interesting parallels that it shows with the notion of extraposition proposed for one of the relationships between rhetorical structure and document structure described by Bouayad-Agha et al. (2000). In the document structure case, however, we see realizations involving indented itemized lists and similar textual variations. In the layout structure case described here, we find distinct visual blocks created on the page, *each of which* can then be subject to its own linear text flow. The typographical variation across distinct layout blocks is also considerably broader than that typically found within any text flow. Both the spatial and the typographic properties of the resulting layout structure elements therefore motivate a distinct treatment.

4 Two Page Examples and some Variations

The three 'spatialization' strategies of the previous section may occur at different places in one and the same RST tree, resulting in a very large number of possibilities. In order to experiment with the mechanisms proposed, we developed an XSLT-based prototype implementation of the process. The resulting mechanism creates from a given RST structure a layout structure formed out of layout units, attributing them with certain typographical features and relative positions on the page. The typographical features and page positions are specified in terms of the layout model described above. The particular values for the elements are taken from our empirical studies. Crucially, therefore, we argue that this breaking of RST relations and assignment to layout chunks is a *genre dependent process*. It is an ongoing goal to ascertain which constraints on the decomposition process can be allocated to genre considerations, which to canvas constraints, and which still remain free.

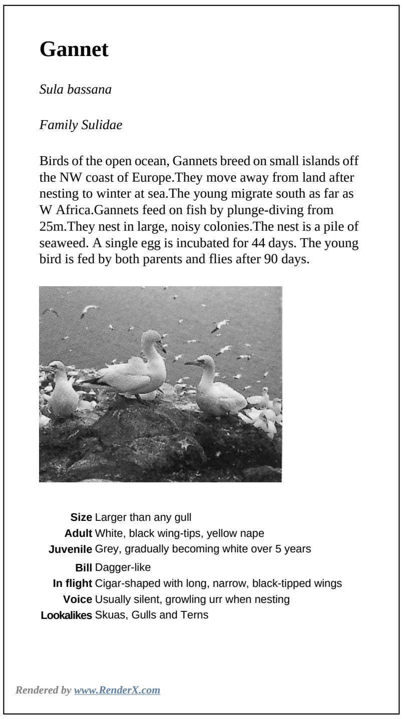
The prototype implementation has been described in more detail in Henschel et al. (2002) and so we will not repeat this here. Two central details of the process are that (a) conditional breaks (cut or emphasis) are potentially available for all satellite arcs and (b) emphasis cuts are potentially available for the nuclei of multinuclear relations. Whether or not a cut is performed depends on *break conditions* defined over any of the GeM description layers. These break conditions—conditions which trigger the start of an extra layout element and break the traditional sequential text arrangement—can draw on several different sources, including:

1. the semantic content of the satellite,
2. the mode in which the satellite should be expressed,
3. the type of RST relation,
4. structural properties of the RST structure.

The examples to follow work with all of these four types of information. We have extracted the most apparent break conditions for bird guides and telephone manuals from the material in our corpus by manual inspection, although it would be useful to consider methods for deriving break conditions automatically in the future.

Whenever layout elements are created by cutting the rhetorical structure, it is necessary to consider their spatial and typographic properties. Both are taken from a genre specification that may be more or less refined. Detailed genre specifications may resemble descriptions of 'house style' or traditional style sheets, whereas less detailed genre specifications characterise the kinds of divisions that should be distinguished in a document without providing exact information about how that is achieved. The precise nature of such genre specifications requires considerably more work in order to ascertain the best distribution of effort: that is, much can already be done by standard rendering engines and we do not intend to reduplicate that work but to pinpoint where document-appropriate flexibility may best be added.

For example, one generation strategy would be to fix an area model on the basis of a genre (for example, a particular page model selected within a document) and to allocate quite specific rhetorical elements to the layout elements available. If those rhetorical elements were also labelled functionally in terms of a genre-specific document structure, then standard rendering mechanisms such as Cascaded Style Sheets[2] or XML Style Language Formatting objects[3] could (or may in the near future) go a long way to producing the final result. For the examples discussed here, we provide minimal genre specification in order to focus on the variability that is supported. We assume a sequential layout as the default layout structure and specify some minimal, rather generic break conditions. The prototype then produces an XSL-FO formatting object document for subsequent rendering. For the examples in this paper we have used the RenderX system.[4]



Gannet

Sula bassana

Family Sulidae

Birds of the open ocean, Gannets breed on small islands off the NW coast of Europe.They move away from land after nesting to winter at sea.The young migrate south as far as W Africa.Gannets feed on fish by plunge-diving from 25m.They nest in large, noisy colonies.The nest is a pile of seaweed. A single egg is incubated for 44 days. The young bird is fed by both parents and flies after 90 days.

Size Larger than any gull
Adult White, black wing-tips, yellow nape
Juvenile Grey, gradually becoming white over 5 years
Bill Dagger-like
In flight Cigar-shaped with long, narrow, black-tipped wings
Voice Usually silent, growling urr when nesting
Lookalikes Skuas, Gulls and Terns

Rendered by www.RenderX.com

Figure 2: Generated bird page

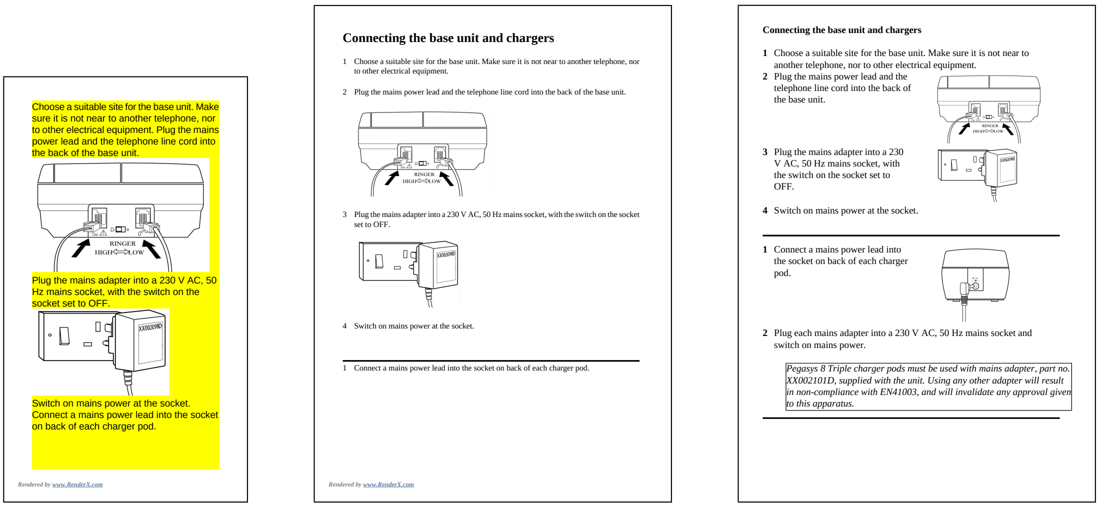
The first example is shown in Figure 2, illustrating the generation of a bird guide page. Such guides consist of a list of similarly appearing document pieces, each of which is dedicated to one particular bird. The generated page closely approximates the original on which it is modelled. Within this genre, we observe that the Latin name of the bird in focus and the information about its family are typically separated layout units. In addition, each bird page has at least one central picture showing the bird (realized in older books as a line-drawing, in more recent ones as a photograph); this graphical layout element is also separated from the textual information. In 'one-bird-per-page' books, the information to be presented in textual form is usually split into two layout units—one represented as linear text, the other as an itemized list. These two layout elements typically have differing typographic realizations, indicating the nucleus-satellite distinction. Typically, each bird page has its own title. Even in a superficially linear (vertical) page such as this, we can readily see that there are actually several distinct layout elements, each with its own distinctive typographical renderings and substructures. These layout elements are aligned across the page *spatially* to enforce Gestalt notions of 'good continuation' and alignment. Only with this kind of modelling can we track what occurs when we move *across* genres to documents where the page layout is constructed differently.

We see this in more detail in the second example, where we also briefly examine some of the variation that can be produced. The rhetorical structure and content of the input corresponds to the information necessary for generating a page from a telephone manual instructing the user how to install a new telephone device. If we apply no genre specific constraints and allow our default sequential layout to apply, then a page with a single layout element is rendered as shown in Figure 3(a), depicted here with the text highlighted according to layout area as used in Figure 1 above. If we then map the rhetorical structure expressing the sequence of instructions to to produce a layout structure consisting of a spatially aligned array of blocks including explicit substructure for labelling, we arrive at the page shown in Figure 3(b). Finally, if we allow page generation to proceed identically to the bird example, the distinct rhetorical structure and genre-specific break conditions gives rise to the page shown in Figure 3(c). In this page, the graphical diagrams have been broken out into their own layout chunks, appearing alongside the layout chunks of the enumeration, and a final elaboration has been broken out in a display box by an emphasis cut.

[2] CSS: `http://www.w3.org/Style/CSS`, particularly the 'paged media' and multicolumn layout additions planned for CSS3.

[3] XSL-FO: `http://www.w3.org/TR/xsl`, particularly the planned flowmap additions.

[4] `http://www.renderx.com`



Choose a suitable site for the base unit. Make sure it is not near to another telephone, nor to other electrical equipment. Plug the mains power lead and the telephone line cord into the back of the base unit.

Plug the mains adapter into a 230 V AC, 50 Hz mains socket, with the switch on the socket set to OFF.

Switch on mains power at the socket. Connect a mains power lead into the socket on back of each charger pod.

Rendered by www.RenderX.com

(a) Default Sequential

Connecting the base unit and chargers

1 Choose a suitable site for the base unit. Make sure it is not near to another telephone, nor to other electrical equipment.

2 Plug the mains power lead and the telephone line cord into the back of the base unit.

3 Plug the mains adapter into a 230 V AC, 50 Hz mains socket, with the switch on the socket set to OFF.

4 Switch on mains power at the socket.

1 Connect a mains power lead into the socket on back of each charger pod.

Rendered by www.RenderX.com

(b) Sequential + Itemized

Connecting the base unit and chargers

1 Choose a suitable site for the base unit. Make sure it is not near to another telephone, nor to other electrical equipment.

2 Plug the mains power lead and the telephone line cord into the back of the base unit.

3 Plug the mains adapter into a 230 V AC, 50 Hz mains socket, with the switch on the socket set to OFF.

4 Switch on mains power at the socket.

1 Connect a mains power lead into the socket on back of each charger pod.

2 Plug each mains adapter into a 230 V AC, 50 Hz mains socket and switch on mains power.

Pegasys 8 Triple charger pods must be used with mains adapter, part no. XX002101D, supplied with the unit. Using any other adapter will result in non-compliance with EN41003, and will invalidate any approval given to this apparatus.

(c) Genre-restricted

Figure 3: Variations on a theme: successive addition of genre-motivated constraints

5 Discussion and conclusions: technical and theoretical issues

Although we have shown that it is possible to begin to move constraints obtained from an empirical analysis of multimodal pages into a page generation process that is not restricted to a largely linear text flow, there are still many problems to be addressed of both a technical and theoretical nature.

The experimental prototype using XSLT and XSL-FO is inherently restricted at present because none of the currently available layout specification languages provides the capabilities assumed by the GeM layout layer. In order to achieve a distribution of information into layout chunks across an entire page, for example, we still have to simulate this by, in effect, creating tables in terms of formatting objects. The working drafts of the planned next versions of both the CSS and XSL-FO will approach the required capabilities more closely, but it is still not clear whether this will be sufficient. We are considering alternative strategies, such as, for example, producing XML-based input to more sophisticated document processing tools, such as Adobe's InDesign or QuarkXpress, and using these as renderers for more varied page design; initiatives such as the Oasis Open Document Format may also be of assistance here.

It is interesting that it is precisely in this move from textual linearity to spatial-visual layout that the available tools let us down. A description such as Power *et al.*'s document structure appears to provide a good model of what happens *within* flow objects carrying text and is readily supported by the currently available technology. In contrast, the visual approach covered by our layout structure is concerned more with just what flow and other entities will be *selected for colocalisation on the page at all* and so represents a decision far closer to the decisions that are taken by document designers. This will often, as shown for one example in detail in Bateman et al. (2001) and formalised above in our 'cutting' operations, involve separating rhetorically related content areas and allocating them to distinct large-scale layout chunks. *Within* those chunks, we can see the rules of document structure proposed by Power *et al.* operating, forming paragraphs, itemized lists and the like. *Outside* of those chunks, we need a further level of visual-spatial organisation that groups them into a hierarchy of perceivable configurations.

Although this in general serves to confirm the observations of Bouayad-Agha et al. (2000) concerning the relationship between punctuation, itemization, etc. and rhetorical structure, it also goes further to show that the phenomenon is more generally applicable. A general rule seems to be that certain RST subtrees are broken apart from the RST tree and rendered as extra layout chunks. To describe this abstractly, we adopted above the metaphor of *extraposition* suggested by Bouayad-Agha *et al.* but now see this as applying to all levels of presentational structure, both layout and document. Precisely how the two layers may interact or be related is therefore an

Mode of symbolization	Method of configuration						
	pure linear	linear interrupted	list	linear branching	matrix	non-linear directed viewing	non-linear most options open
Verbal/ numerical	1	2	3	4	5	6	7
Pictorial & verbal/ numerical	8	9	10	11	12	13	14
Pictorial	15	16	17	18	19	20	21
Schematic	22	23	24	25	26	27	28

Table 2: Twyman's (1987) characterisation of distinct visual presentational styles

interesting topic for future research.

It is sometimes suggested, for example by Bouayad-Agha (2001), that the genres for which visual organisation is significant may represent something of a minority among naturally occurring documents. However, once our analytic approaches are able to respond to this kind of information, we in fact see that the visual component is rarely absent. Scott and Power (2001), for example, have now suggested that their former view of document structure may need to be extended in the face of textual presentations which employ visual resources, such as tables and certain menu-driven interactions. They come to the conclusion that diagrams and text may not be as distinct as commonly proposed—although this has also been shown concretely in the common text/diagram aggregation algorithms presented in Bateman et al. (1998). Many documents make considerably more flexible use of the visual/spatial resource than simple extensions of a linear model will support, even when dealing with largely textual presentations. And, moreover, this is a very common situation.

The importance of assigning more weight to more systematic accounts of the visual-spatial dimension of document layout can be seen from Twyman's (1987) characterisation of distinct types of visual-graphical page organisations. We reproduce his 'matrix' of possibilities in Table 2. The kind of 'macro-punctuation' organisation captured by document structure is then covered by the linear categories, in particular Twyman's types 1, 2, and 3. If we in addition allow graphical elements into this essentially linear structure, this brings in types 8, 9, and 10. As soon as the spatial possibilities of the page are taken into consideration, however, the types in the columns 'linear branching', 'matrix' and 'non-linear directed viewing' all become relevant also (4–7, 11–14). In our work on GeM, it is precisely the documents of these latter groups that have been our primary focus: the deployment of the page as a two-dimensional viewing space provides for the 'non-linear' component, while the selection of particular typographically signalled layout chunks functions to direct viewing, or to provide the *access structure* in the terms defined in the GeM scheme (Bateman et al., 2004).

Rather than representing an exception to the rule, these classes of documents are fast becoming the norm in all areas where sophisticated layout is deployed: from entertainment magazines to educational text books. It is therefore of considerable importance for multimodal document generation to have appropriate means for studying and producing such genres. Both the approach to layout structure as such that we have presented here and our proposals for incorporating sensitivity to layout structure in the generation process are intended as steps to bring us closer to the generation of such non-linear spatially organised documents.

References

Allen, P., Bateman, J. A., and Delin, J. (1999). Genre and layout in multimodal documents: towards an empirical account. In Power, R. and Scott, D., editors, *Proceedings of the AAAI Fall Symposium on Using Layout for the Generation, Understanding, or Retrieval of Docu-*

ments, number Technical Report FS-99-04, pages 27–34, Cape Cod, Massachusetts. American Association for Artificial Intelligence.

André, E., Finkler, W., Graf, W., Rist, T., Schauder, A., and Wahlster, W. (1993). WIP: the automatic synthesis of multimodal presentations. In Maybury, M. T., editor, *Intelligent Multimedia Interfaces*, pages 75–93. AAAI Press/The MIT Press, Menlo Park (CA), Cambridge (MA), London (England).

Anjewierden, A. (2003). A library of document analysis components. Technical report, Social Science Informatics, University of Amsterdam.

Bateman, J. A., Delin, J. L., and Henschel, R. (2002). A brief introduction to the GEM annotation schema for complex document layout. In Wilcock, G., Ide, N., and Romary, L., editors, *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002) — Post-Conference Workshop of the 19th International Conference on Computational Linguistics (COLING-2002)*, pages 13–20, Academica Sinica, Taipei, Taiwan. Association of Computational Linguistics and Chinese Language Processing, Association of Computational Linguistics and Chinese Language Processing.

Bateman, J. A., Delin, J. L., and Henschel, R. (2004). Multimodality and empiricism: preparing for a corpus-based approach to the study of multimodal meaning-making. In Ventola, E., Charles, C., and Kaltenbacher, M., editors, *Perspectives on Multimodality*, pages 65–87. John Benjamins, Amsterdam.

Bateman, J. A., Kamps, T., Kleinz, J., and Reichenberger, K. (1998). Communicative goal-driven NL generation and data-driven graphics generation: an architectural synthesis for multimedia page generation. In *Proceedings of the 1998 International Workshop on Natural Language Generation*, pages 8–17. Niagara-on-the-Lake, Canada.

Bateman, J. A., Kamps, T., Kleinz, J., and Reichenberger, K. (2001). Constructive text, diagram and layout generation for information presentation: the DArt$_{bio}$ system. *Computational Linguistics*, 27(3):409–449.

Bouayad-Agha, N. (2001). *The role of document structure in text generation.* ITRI Report ITRI-01-24, University of Brighton, Information Technology Research Institute, Brighton, UK.

Bouayad-Agha, N., Power, R., and Scott, D. (2000). Can text structure be incompatible with rhetorical structure? In *Proceedings of the International Natural Language Generation Conference (INLG-2000)*, pages 194–200, Mitzpe Ramon, Israel.

de Carolis, B., de Rosis, F., and Pizzutilo, S. (1997). Generating user-adapted hypermedia from discourse plans. In *Proceedings of Fifth Congress of the Italian Association for Artificial Intelligence (AIIA'97)*, pages 334–345, Rome.

Delin, J., Bateman, J. A., and Allen, P. (2002/3). A model of genre in document layout. *Information Design Journal*, 11(1):54–66.

Eglin, V. and Bres, S. (2004). Analysis and interpretation of visual saliency for document functional labeling. *International Journal on Document Analysis and Recognition (IJDAR)*, 7:28–43.

Feiner, S. K. (1988). A grid-based approach to automating display layout. In *Proceedings of the Graphics Interface*, pages 192–197, Los Angeles, CA. Morgan Kaufman.

Goldfarb, C. F., editor (1990). *The SGML Handbook.* Clarendon Press, Oxford.

Graf, W. H. (1995). The constraint-based layout framework laylab and its applications. In *Proceedings of ACM Workshop on Effective Abstractions in Multimedia, Layout and Interaction*, San Francisco, California. ACM.

Henschel, R., Bateman, J., and Delin, J. (2002). Automatic genre-driven layout generation. In *Proceedings of the 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, pages 51–58, University of the Saarland, Saarbrücken.

Holsanova, J., Rahm, H., and Holmqvist, K. (2006). Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements. *Visual Communication*, 5(1):65–93.

Hovy, E. H. and Arens, Y. (1991). Automatic generation of formatted text. In *Proceedings of the 8th. Conference of the American Association for Artifical Intelligence*, pages 92–96, Anaheim, California.

Kruijff, G.-J., Teich, E., Bateman, J. A., Kruijff-Korbayová, I., Skoumalová, H., Sharoff, S., Sokolova, L., Hartley, T., Staykova, K., and Hana, J. (2000). A multilingual system for text generation in three Slavic languages. In *Proceedings of the 18th. International Conference on Computational Linguistics (COLING'2000)*, pages 474–480, Saarbrücken, Germany.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Nunberg, G. (1990). *The Linguistics of Punctuation.* Number 18 in CSLI Lecture Notes. Center for the Study of Language and Information, Stanford.

Power, R. (2000). Mapping rhetorical structures to text structures by constraint satisfaction. Technical Report ITRI-00-01, ITRI, University of Brighton.

Power, R., Scott, D., and Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, 29(2):211–260.

Reichenberger, K., Rondhuis, K., Kleinz, J., and Bateman, J. A. (1995). Effective presentation of information through page layout: a linguistically-based approach. In *Proceedings of ACM Workshop on Effective Abstractions in Multimedia, Layout and Interaction*, San Francisco, California. ACM.

Scott, D. and Power, R. (2001). Generating textual diagrams and diagrammatic text. In Bunt, H. and Beun, R.-J., editors, *Cooperative Multimodal Communication*, Lecture Notes in Artificial Intelligence. Springer, Berlin.

Sefton, P. M. (1990). Making plans for Nigel (or defining interfaces between computational representations of linguistic structure and output systems: Adding intonation, punctuation and typography systems to the PENMAN system). Technical report, Linguistic Department, University of Sydney, Sydney, Australia. Batchelor's Honours Thesis.

Summers, K. (1995). Toward a taxonomy of logical document structures. In *Electronic Publishing and the Information Superhighway: Proceedings of the Dartmouth Institute for Advanced Graduate Studies (DAGS '95)*, pages 124–133, Boston, MA.

Summers, K. (1998). *Automatic discovery of logical document structure.* PhD thesis, Cornell University, Ithaca, New York.

Thompson, H. S. and McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97.*

Twyman, M. (1987). A schema for the study of graphic language. In Boyd-Barrett, O. and P., B., editors, *Media, Knowledge and Power*, pages 201–225. Croom Helm, London. Version reprinted from: Kolers, P. A., Wrolstad, M. E., and Bourna, H., editors, *Processing of Visible Language*, volume 1, pages 117–150. Plenum, New York and London.

Non-Localized, Interactive Multimodal Direction Giving

Charles Callaway
University of Edinburgh
2 Buccleuch Place
Edinburgh, UK EH8 9LW
ccallawa@inf.ed.ac.uk

Abstract

Many modalities have been utilized for the presentation of route directions to users with mobile devices, such as 3D graphics, 2D maps, 2D diagrams, as well as written and spoken text. Most such direction giving systems either use localization technologies and can thus be interactive, or are non localized and as a consequence non-interactive. We describe the iterative design of a prototype for non-localized yet interactive mobile indoor direction giving that involved varying degrees of natural language generation and 2D graphics to give directions. Multimodality assists in the presentation of the route instructions themselves, but even more so when helping users recover their location if they get lost. A detailed but informal experiment with our prototype helped determine which characteristics of the graphical user interface, specifically its graphics and text modalities, are useful for giving directions in a helpful way.

1 Introduction

Most current interactive direction giving systems require the use of localization hardware, such as wireless beacons, GPS, or infrared sensors (Cheverst *et al.* 2000; Baus *et al.* 2002; Müller 2002; Kray *et al.* 2003), to successfully navigate to the selected destination. These systems can use localization to plan or replan a trip to a new destination by using the current position as the starting point. They work best in outdoor environments and typically prefer 2D graphical representations over natural language output such as speech or text. Commercialized car navigation systems often use text-to-speech synthesis, but are restricted to street names and very simple relationships between them such as "turn left at". By and large these systems use multimodality for ease-of-use (*e.g.*, hands free) rather than improving the quality of the directions themselves.

Other methods of providing directions without localization hardware, such as commercial web services that provide sequences of driving instructions (*e.g.*, MapQuest or YahooMaps) do not provide true interactivity (Geldof and Dale 2002). For example, a driver with a printout of directions to some destination who becomes lost must return to a previously known point in those directions and then continue with the original instructions, or else be able to discern their current location and reconnect to the website to generate new directions. Thus the current state of the art poses a tradeoff between the expense and restrictions of localization hardware against the usefulness of interactivity.

Direction giving systems can be further characterized according to what degree they employ techniques in natural language generation. A major desire of such systems is to produce directions comparable to those of humans in a particular situation (Dale *et al.* 2002). Thus a number of corpus studies have taken place to determine the nature of the syntax used (Stoia *et al.* 2006), the characteristics of situations where particular sentence types are used (Look *et al.* 2005), as well as what types of gestures are likely to be used (Striegnitz *et al.* 2005). However, these studies thus attempt to reproduce existing human direction-giving behavior, and to compete against the strengths of localization systems which have access to detailed maps and existing well-placed descriptors like street signs. No effort has been directed to determining new scenarios that play to the strengths of natural language generation and multimodality and which highlight weaknesses in existing localization systems.

Spatial representation is a further area of research, especially for systems that are interactive and are lacking in localization, since localization methods such as GPS often allow for simpler Cartesian coordinate systems. Indoor direction giving in large buildings on the other hand, with its inherently large number of obstacles, resistance to simple hardware localization like GPS, and unnamed corridors, is less amenable to such approaches. Indoor direction-giving systems thus tend to make extensive use of landmarks and explicitly hand-mark paths which the user must be guided along, whereas outdoor graphics-based localized systems need only show the user's current position and destination on a 2D map, allowing the user to judge the best course of action to follow. Localization also by its nature precludes users losing their way, while nonlocalized text based systems, interactive or not, are not able to offer advice on *location recovery*. Indeed, because corpus studies have yet to address what behaviors people use when they do get lost, they would not even know how to offer such advice.

We have conducted an informal experiment in a new type of direction giving scenario: non-localized, interactive route finding where the user can assist in the explicit localization of their own position both indoors and outdoors. The lack of hardware localization implies that we are unable to collect second-by-second feedback on the user's current position, but must instead resort to cooperative user interaction via a mobile interface designed for this purpose. This method may not be usable in all scenarios, such as those where the user is constrained by the amount of time available or when driving, or even be optimal compared to hardware-based localization when circumstances favor that approach. The work presented here is aimed at exploring how to improve multimodal and language-based route descriptions in indoor pedestrian settings using simple mobile tecnology such as PDAs or cellphones.

2 Producing directions

Although much current research in direction giving is devoted to graphical representations (Baus *et al.* 2002; Müller 2002) and multimodality that consists of pen or voice input with output on PDAs, especially for GPS-based outdoor systems, a significant subset uses natural language generation (NLG) techniques to automatically produce text usable for giving directions. Thus for instance (Fraczak *et al.* 1998) focused on the relative importance of certain path segments and how they affected both the content and linguistic form of directions. Additionally, the Coral project (Geldof and Dale 2002) used a non-localized PDA interfaced to the Australian WhereIs.com driving directions website to create individual textual descriptions for each path segment. They also described helpful constraints for presenting textual directions on mobile devices. None of these approaches relies on multimodality, with the first group looking only at graphics and the second using generated language only.

Many of the types of directions given have fallen into one of four categories: acknowledgements of current position, turning or other movement, reorienting while remaining in position, and giving references based on landmarks. Strategies that lack localization technology tend to make significant use of landmarks, and there are differences in the treatment of landmarks depending on whether a direction giving system is text- or graphics-based. Landmarks are the focus of much current research in direction giving (Burnett 2000; Raubal and Winter 2002) and have significant effects on the text that should be generated in a given situation. Little attention has been paid to indoor landmarks (since text-based NLG systems are almost exclusively for outdoor directions) and the effect of culture and other factors on their salience.

Some direction giving systems produce natural language output but do not use NLG techniques to do so. Lair (Look *et al.* 2005) is intended as an experiment in modeling spatial representations. Its world consists of *places*, which have a predetermined list of paths they lie on, *paths*, which are ordered lists of places, a list of the geometrical relationship between all paths at one place, a list of other places visible from the current place, containment relations of one place within another, and a list of functions that can be performed at the current place. Lair guides visitors at the MIT CSAIL lab, using A* search to find a path from one location to another. They collected a corpus of human directions in the same space, learning how to collapse sequences of waypoints when possible (a form of high-level revision).

(Stoia *et al.* 2006) describe interactive directions in indoor spaces considering the visibility of and distance to reference points. However their focus was mainly on how to correctly generate noun phrases in given situations compared to human directions. To this end they created several methodologies for evaluation, including 3D reconstructions. Corpus collection and annotation played a large role in their project, leading to the gathering of situational features which have an impact on the NPs produced.

(Striegnitz *et al.* 2005) describes the beginning of a dialogue direction giving system that is also based on corpora of humans giving directions. However here the focus is on the gestures that people use together with verbal directions, especially the basis for the dialogue itself and the knowledge representation needed to generate text and gestures indicating the location of landmarks. Thus it is multimodal, but as a way to understand human direction giving rather than show how multimodal presentation can increase success. Like other systems, it requires maps annotated with paths that a user should take. The system keeps track of a person's and embodied agent's position and orientation, allowing it to produce gestures that accurately direct the user's attention to salient landmarks.

3 Important factors affecting direction giving

Many factors affect the style of interaction between a user and a direction giving system that uses language. Among them are:

- **Naturalness of language:** Deep natural language generation, which uses sophisticated linguistic operations such as revision and referring expression, can make interactions with a PDA system very close to the types of written or spoken directions given by people (Dale *et al.* 2002). This allows for linguistic techniques that pack a large amount of relevant information into a very short amount of space and extends beyond lexical and grammatical considerations. For instance, path unit segmentation should be performed that will mimic the types of segmentations used by human guides (Geldof and Dale 2002), and properties of landmarks should be referred to by features that people find most salient.

- **Multimodal Output:** The most significant problems for multimodality involving text generation are synchronization and media selection. If 2D or 3D graphics are used in conjunction with generated text the timing (Towns *et al.* 1998) and content (Foster and White 2006) in each must be synchronized and not contradictory. Main content must also be allocated to a particular media with supporting content in secondary media, as users are quick to notice redundancy. Overlapping information can be beneficial when it ensures that information is received by the user, or harmful in that it may cause confusion. Finally, presenting graphics and text together on mobile screens is often challenging due to system maintenance (e.g., one developer changes red arrows in the graphics component to green, while the text continues to mention red arrows) and space constraints for displaying text, although using voice output to solve this also presents its own set of problems.

- **Interactivity:** Systems can be (1) completely interactive, ranging from dialogue systems that allow the user to change their minds about what to do at any point to location-aware systems that constantly update position relative to the destination, (2) semi-interactive where changing one's mind involves reinitializing a system with new start and end points as well as other parameters and then replanning, and (3) completely non-interactive, such as printed driving instructions from a web service where changing one's mind requires returning to a computer to print new directions. Each type requires different types of knowledge and little re-use can be expected for different types.

- **Landmarks** are especially useful when there is no automatic localization method (Burnett 2000). As suggested in (Raubal and Winter 2002), visual salience for landmarks is highly important, in order to distinguish them from other nearby distractors, and thus lessons learned from Referring Expression Generation can be very important when determining what can be a landmark and how to describe it (for instance, brick vs. steel is better than 20 meters tall vs. 25 meters tall). However, computational methods for automatically determining visual salience are still in their infancy, and thus all landmarks across all types of direction giving systems are currently hand-annotated. In addition, landmarks that function well when driving might be inappropriate for pedestrians and vice versa, landmarks that lie exactly along a right may be the worse choice if they are not distinguishable (e.g., parking meters), and outdoor landmarks may have different types of distinguishing properties than indoor landmarks as well as requiring different types of language to describe them.

- **Cultural Differences for Landmarks:** Part of the visual saliency of landmarks is due to the user's individual perspective. Objects which appear frequently in one culture may have their landmark value diminished for users of that culture due to their ubiquity, but for a user from a different culture, that object may be novel and thus be an optimal landmark. For instance, in the experiment described below, Americans and Italians differed significantly on whether an umbrella holder could be considered an indoor landmark. This knowledge is especially useful in domains such as tourism, where direction giving is highly likely to be a necessity and the backgroundy of each user is highly variable.

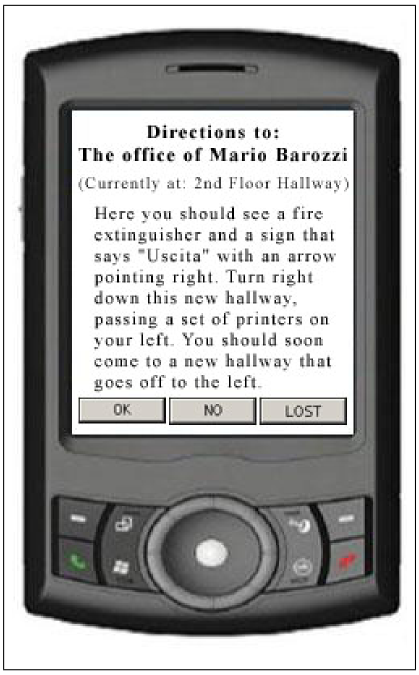
- **Indoor vs. Outdoor:** Landmarks are typically very scarce in indoor environments, while there are often too many to choose from in outdoor environments. But landmarks are even more important indoors due to the lack of other localizable features that can be used such as street signs. Most direction giving systems are intended to be used exclusively in one environment or the other, with the notable exception of (Baus *et al.* 2002), which can switch between the two, but is dependent on the existence of special infrared beacon systems. Indoor navigation by GPS is often useless in buildings, especially many tourist destinations in Europe that have thick rock walls which block satellite and wireless signals. Additionally, standard GIS databases cover only outdoor areas. Thus one of the aims of direction giving research is finding models that work seamlessly both in and outdoors using only hardware that potential users already carry with them.

- **Spatial Representation:** Systems often employ a Cartesian representation, especially in graphical direction giving systems, either with or without localization. An alternate choice used more often in indoor systems is a discrete representation of objects and positions in the world such as waypoints and paths, along with coded relationships between them that allow for a wider range of directions to be generated automatically by, for instance, solving a path problem with standard AI search and then generating text by traversing the nodes of the search path. But spatial representations would also be ideal as an attachment point to external databases that contain for instance landmarks along with their properties.

- **Getting Lost:** When hardware localization is not involved, users can sometimes become lost and require help in getting back on track. For instance, while following driving directions a user may have stopped for a drink or encountered someone, causing them deviate from the planned path. Alternately, a user might actively realize several planned steps later that they have followed directions incorrectly. In both cases, the system needs to help users locate themselves before replanning a new path. Here multimodality can help by, for example, showing combinations of pictures of landmarks and interactive natural language dialogue allowing the user to quickly reorient themselves and plan a new route if necessary. Below, we discuss several potential scenarios and suggestions that use multimodality and natural language generation together to solve this relocalization problem.

4 Basic scenario

Our direction giving scenario examines indoor locations, since outdoor direction giving is currently well-covered by GPS services, all of our potential users were readily familiar with the indoor



Directions to:
The office of Mario Barozzi

(Currently at: 2nd Floor Hallway)

Here you should see a fire extinguisher and a sign that says "Uscita" with an arrow pointing right. Turn right down this new hallway, passing a set of printers on your left. You should soon come to a new hallway that goes off to the left.

Figure 1: Text-only interactive directions

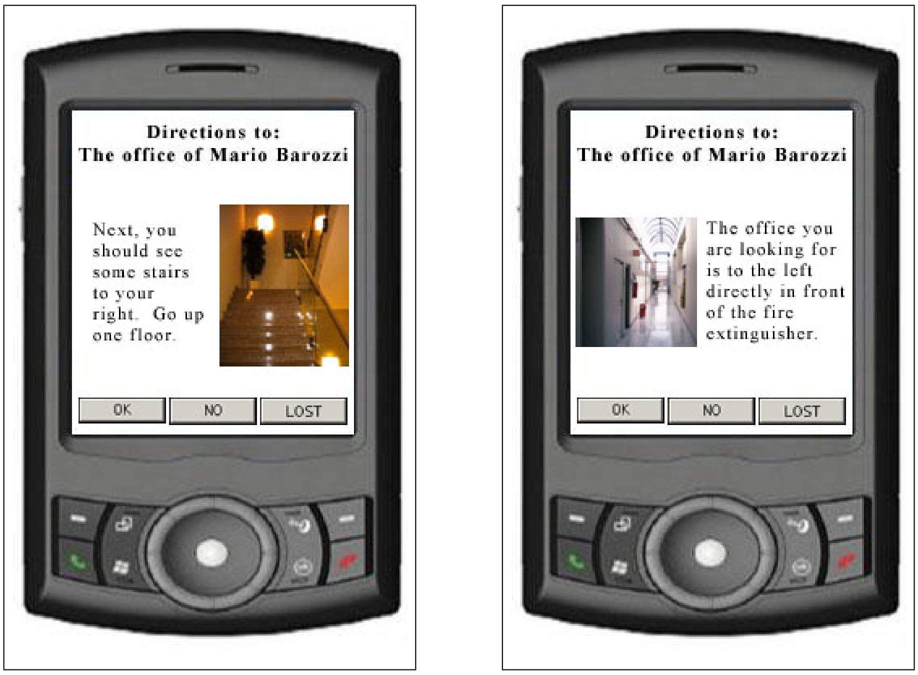
location we chose, following users on meaningful pedestrian-oriented outdoor trips is much more resource intensive, and the use of indoor landmarks is less well studied than those outdoors, allowing for interesting studies in multimodality.

For our scenarios we utilized a three-story office building where none of the participants frequently venture to any parts of the building besides the entrance and their work location. We thus created a prototype, initially text-only though interactive (Figure 1), with the purpose of having users begin at the entrance to the building and finding their way to a destination in the building.

A spatial representation of the building was created with both *waypoints*, junctions of paths where users can make choices, and property-based formal descriptions of visual elements the user might see. For experimenting with multimodality, we used a digital camera to take pictures of the surroundings and the landmarks of paths and added the most salient among them to a database containing descriptions of their properties, such as color, size, orientation, material, etc.

The major distinction between interactive and non-interactive systems is that with the former users can make choices and choice points. Thus the standard method of presenting interactive directions is to deliver a chunk of instructions between two waypoints called a *path segment*. When the user arrives at the waypoint, the next batch of instructions can be delivered. Without localization, it is up to the user to verify that they have correctly arrived at the new waypoint. Thus, arriving at each new waypoint is an opportunity for a new multimodal presentation, as are declarations by the user that they have either returned to or are unable to return to a previous waypoint.

Standard indoor direction giving does not require new knowledge in multimodal presentation strategies, as these techniques are both simple and already known. To ensure we would have enough material to study the use of multimodality, we allowed our users to follow approximately one third to one half of their path before simulating their “being lost” by taking them to another area of the building without access to the PDA until they arrived, and then only allowing them to reorient themselves by using the PDA again. Our goal was to determine which presentation strategies work best in this type of situation and learn about how to design future formal experiments. Our assumption was that mixed graphics and text would work best for this task.



Directions to:
The office of Mario Barozzi

Next, you should see some stairs to your right. Go up one floor.

Directions to:
The office of Mario Barozzi

The office you are looking for is to the left directly in front of the fire extinguisher.

Figure 2: Two path segments in our prototype evaluation

5 Prototype experiments

Because constraints on the user interface and use of multimodality are not well known, we set out to perform an iterative, informal study that would allow us to discover what rules govern the combined use of text and graphics in a non-localized, interactive mobile environment. The interaction scenario described above involves the use of a standard PDA with space to display text as well as three buttons for the user to indicate whether she has understood or not, or thinks she has become lost.

Our initial constraints on text generation and display were taken from common sense as well as the literature: (1) there is a small amount of screen space to display text, (2) any and all of the display space is available for a combination of text and graphics, (3) greater descriptive detail about landmarks was better than a minimal amount of detail, (4) options such as screen scrolling that hinder text readability and require continuous user intervention are not desirable, (5) users expect to be able to review previous directions, especially if they think they are lost, and (6) graphics should be better than text for relocalizing lost users. Additionally, we were interested in the effectiveness of the instructions themselves, especially in forming hypotheses about differences between text or graphics alone versus text with graphics.

In our formative study, a random destination in a building (as described above) was assigned to a user who rarely entered that part of the building. Our prototype presented directions one path segment per screen, as shown in Figure 2, where the directions for that path segment would remain onscreen until one of the buttons was selected. After the system's presentation of each segment, the user gave initial impressions to the evaluator, who recorded them and asked only superficial clarification questions. User cooperation was assumed throughout the experiment, and the user comments were used to refine the interface in subsequent trials. A summary of these comments, especially those regarding multimodality, are given in the following section.

The presentation of text directions and 2D graphics were canned to guarantee rapid prototyping, as the focus of the study was on the perceived effectiveness of combining text and graphics rather than correctness or effect of language. In this manner we could carefully control as many non-interface variables as possible, such as synchronization of text presentation with real-time user movements. The only modifications to the text presented to the user across all trials were to adjust the amount of detail describing relevant landmarks.

We also experimented several multimodal methods to allow users to recover their location should they become lost. This involved five gradations ranging from a text-only description where the user had to select the closest match from a short list of paragraphs describing locations nearest their last correct waypoint response, to interactively presenting them simultaneously with 2D images and very short descriptions of salient landmarks, to a graphics-only version where actual pictures of parts of the building were shown. We tested this by purposefully "moving" users to nearby locations (simulating their incorrect application of earlier directions), forcing them to reorient themselves and gauging their responses to the various types of single and multimodal presentations.

6 Lessons learned on multimodality

The principle lessons we learned about presenting textual descriptions that violated the hypotheses described above fell into the following categories:

- **Length:** Users preferred shorter instructions to longer ones, as long as they were still unambiguous. A more important constraint than preference, however, was utility. For instance, some of the waypoints (which are ad-hoc markers) were "physically" very close in our spatial model, especially near large intersections. When users were given even moderately long instructions (half a PDA screen of text), they frequently walked past even more waypoints as they continued to read the directions, and were thus unable to follow the original instructions when they finished reading because they no longer applied to their new location. One possible solution is to base restrictions on text length to the length of the current and upcoming path segments (if dynamic text length restrictions are available). Alternately, spatial representations can be carefully controlled when they are created to keep a minimum distance between waypoints.

- **Content:** As can be inferred from Grice's maxims (Grice 1975) which are familiar to NLG researchers, users dislike the inclusion of irrelevant information, especially for instance nearby landmarks which are not in the direction of the next path segment, indications of alternate but acceptable paths, or even mentions of places to avoid (*e.g.*, "Don't go down that hallway." Thus landmarks that have low salience should not be mentioned (although salience is a dynamic feature, so a landmark that is not salient in one moment may become so in the next). Grice's maxims have not yet been interpreted from the perspective of multimodality, but it should be noted that we never received unfavorable comments when the same content was duplicated across both the text and graphics modes.

- **Treatment of landmarks:** Landmarks with high salience (and that should therefore be included in the path segment presentation) that are located on the far side of rooms or hallways from the waypoint will often appear very small in 2D images such as pictures, and should therefore be allocated additional space for textual description, and even more helpfully with an accompanying highlight effect. In general, indoor and outdoor landmarks may impose different perceptual constraints, in addition to salience.

- **Emphasizing the textual dimension:** Many suggestions were given by the participants to improve the readability of the text, perhaps indicating that properly exploiting NLG's potential for improvement could have a greater effect than trying to use equal resources on improving graphics. Suggestions included color highlights of the most relevant portions of the text, for example with imperatives such as "go straight ahead" or "turn right" marked in green as opposed to achieving the same aim by adding more descriptive text about landmarks.

  As we had expected, requiring scrolling in the interface because of excessive amounts of text was found to be annoying to the users. The principle HCI concern behind this is that scrolling requires constant effort on the part of the user to maintain their "current reading position" in the text. Fitting all of the text into a single screen, regardless of whether there were additional graphics or not, was unanimously desirable.

- **Emphasizing the graphical dimension:** An initial attempt to use the graphical directional arrows from outdoor GPS interfaces (Baus *et al.* 2002) superimposed on our 2D picture of a room rather than on a map was not successful as we could not update the user's orientation in real time. In general, subjects often evaluate a waypoint having many choice points by turning their bodies as they looked in each direction. This eliminated the usefulness of the arrow as for most of their choices they were facing the wrong direction.

  Separately, the graphics-only condition of the location recovery gradation presented at the end of the previous section (*i.e.*, only show pictures of potential locations) seems to work only with very highly salient landmarks, as less-salient landmarks get washed out by the surrounding clutter. This leads to a possible mode-selection criterion where certain extremes known to be effective can be annotated and presented only in a particular mode (here, graphics) while in other cases they can be presented as combinations of text and graphics.

- **Quasi-multimodal interfaces:** As with the scrolling problem mentioned above, forcing users to discern between graphics and text together in a small space leads to difficulties. For instance, with our intermediate multimodal relocation strategy, we found that typically the entire screen space was needed in order for users to be able to identify a landmark or else no distinguishing detail would be visible. We had much more success with a mode-swapping interface whereby the image was shown on the entire screen area except for the buttons shown in Figure 2. If the user was still unsure about identifying the landmark, a touch anywhere on that image would replace the entire image with the entire textual description. As the swap is quite fast and the text stays in place when it is toggled back and forth with the screen, users found it highly desirable compared to text scrolling.

7 Pointers to future evaluations

We believe this informal study, although limited, raises a number of concrete suggestions for more formal studies in the future. For instance, the frequent requirement that landmarks be highly visually salient before predominantly 2D graphics modes are selected begs the question, how do we determine the salience of most landmarks? For instance, would it be useful to conduct an experiment where subjects are individually shown a series of landmarks and a majority "vote" is then taken to determine an overall salience level for that landmark? Given enough instances and opinions, could we learn a saliency determination algorithm?

Given that arrows superimposed over pictures of locations had limited utility, would other combined uses of graphics and images suffer similarly? Would some more intelligent combinations of them succeed where otherwise graphics and text combined would be best? In general, we did not consider the use of audio, animation, or other modalities for direction giving.

We hypothesize that increasing the amount of property-based knowledge about landmarks would lead to more advantageous use of existing referring expression generation algorithms and thus to better synchronization and "controlled overlap" between text and graphics. How should this be tested? Is task effectiveness an appropriate way to evaluate multimodal capability? What about user satisfaction as the dialogue systems community uses? A combination of the two?

Related work carried out at the same time in the Peach project at Irst in Italy used PDAs as intermediaries to present personalized and dynamically generated documentary films (Callaway *et al.* 2005) that included automatic zooming into individual scene elements and other cinematic transitions in order to bring out relevant details that would otherwise be too small to see on the screen. Multimodality greatly improves the usability of small devices such as PDAs that have limited capability in individual modalities.

Acknowledgements: Elena Not and Alessandra Novello of ITC-irst participated in many discussions on direction giving and how to achieve it without localization.

References

[Baus *et al.* 2002] Baus, J., Krüger, A., and Wahlster, W. A resource-adaptive mobile navigation system. In the *Proceedings of the International Conference on Intelligent User Interfaces*, ACM Press.

[Burnett 2000] Burnett, G.E. Turn right at the traffic lights: The requirement for landmarks in vehicle navigation systems. *The Journal of Navigation*, 53(**3**), pages 499–510.

[Callaway *et al.* 2005] C. Callaway, E. Not, A. Novello, C. Rocchi, O. Stock and M. Zancanaro. Automatic cinematography and multilingual NLG for generating video documentaries. *Artificial Intelligence*, 165(**1**), pages 57–89.

[Cheverst *et al.* 2000] Cheverst, K., Davies, N., Mitchell, K., Friday, A., and Efstratiou, C. Developing a Context-aware Electronic Tourist Guide: Some Issues and Experiences. In the *Proceedings of CHI2000*, The Netherlands, pages 17–24, April.

[Dale *et al.* 2002] Dale, R., Geldof, S. and Prost, J.P. Generating more natural route descriptions. In the *Proceedings of the 2002 Australasian Natural Language Processing Workshop*, pages 41–48, Canberra, Australia, December.

[Dale *et al.* 2003] Dale, R., Geldof, S. and Prost, J.P. Using natural language generation for navigational assistance. In the *Proceedings of the 2003 Australasian Natural Language Processing Workshop*, pages 35–44, Australia.

[Foster and White 2006] Mary Ellen Foster and Michael White. Assessing the Impact of Adaptive Generation in the COMIC Multimodal Dialogue System. In the *IJCAI Workshop on Knowledge and Representation in Practical Dialogue Systems*, Edinburgh, July.

[Fraczak *et al.* 1998] Fraczak, L., Lapalme, G., and Zock, M. Automatic Generation of Subway Directions: Salience Gradation as a Factor for Determining Message and Form. In the *Proceedings of the 9th International Workshop on Natural Language Generation*, Niagara-on-the-Lake, Canada.

[Geldof and Dale 2002] Geldof, S. and Dale, R. Improving route directions on mobile devices. In the *Proceedings of the ISCA workshop on Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany, June.

[Grice 1975] Grice, H.P. Logic and Conversation. In *Syntax and Semantics 3: Speech Acts*, P. Cole and J. Morgan, eds. Pages 41–58, Academic Press, New York.

[Johnstone *et al.* 2002] M. Johnstone, P. Ehlen, S. Bangalore, M. Walker, A. Stent, P. Maloor and S. Whittaker. MATCH: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July 2002, pages 376–383.

[Kray *et al.* 2003] Christian Kray, Katri Laakso, Christian Elting and Volker Coors. Presenting Route Instructions on Mobile Devices. In the *Proceedings of the Eighth International Conference on Intelligent User Interfaces*, pages 117–124, Miami, Florida.

[Look *et al.* 2005] G. Look, B. Kottahachchi, R. Laddaga and H. Shrobe. A location representation for generating descriptive walking directions. In the *Proceedings of the Tenth International Conference on Intelligent User Interfaces*, pages 122–129, San Diego, California.

[Müller 2002] Christian Müller. Multimodal Dialog in a Pedestrian Navigation System. In the *Proceedings of ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, pages 42–44, Kloster Irsee, Germany.

[Raubal and Winter 2002] Raubal, M. and Winter, S. Enriching wayfinding instructions with local landmarks. In the *Proceedings of GISscience 2002*, Lecture Notes in Computer Science, Springer, Berlin.

[Stoia *et al.* 2006] L. Stoia, D. Shockley, D. Byron and E. Fosler-Lussier. Noun phrase generation for situated dialogs. In the *Proceedings of the Fourth International Conference on Natural Language Generation*, Sydney, Australia, July.

[Striegnitz *et al.* 2005] K. Striegnitz, P. Tepper, A. Lovett and J. Cassell. Knowledge representation for generating locating gestures in route directions. In the *Proceedings of the Workshop on Spatial Language and Dialogue*, Delmenhorst, Germany, October.

[Towns *et al.* 1998] Stuart Towns, Charles Callaway and James Lester. Generating Coordinated Natural Language and 3D Animations for Complex Spatial Explanations. In the *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Madison, Wisconsin, pages 112–119, 1998.

Issues for Corpus-Based Multimodal Generation

Mary Ellen Foster
Robotics and Embedded Systems Group
Faculty of Informatics, Technical University of Munich
Boltzmannstraße 3, 85748 Garching, Germany
foster@in.tum.de

Abstract

In recent years, data-driven methods have become increasingly popular in natural language generation. Multimodal generation can also benefit from using corpus data directly; however, there are several issues that arise when using corpora for multimodal generation that do not occur in the unimodal case, and that mean that existing multimodal corpora are often not suitable for being directly used in a generation system.

1 Introduction

In recent years, there has been an increasing amount of interest in the collection, annotation, and use of multimodal corpora—recorded collections of multimodal human behaviour, labelled and annotated for use in tasks such as analysis and summarisation. Another growing field of research is data-driven methods for natural language processing; this began with tasks such as parsing and machine translation, but more recently researchers in natural language generation have begun to take advantage of these data-driven methods as well.

Combining techniques and data from these two fast-growing fields to implement multimodal corpus-driven generation adds extra requirements that do not arise in each of the individual research fields: corpus-based techniques for text generation do not necessarily apply directly to the multimodal case, while general-purpose multimodal corpora are not always suitable for use in a generation system. This paper discusses several of the issues that must be taken into consideration if the two research areas are to be combined.

The paper is structured as follows. In Section 2, we first summarise the state of the art on multimodal corpora and the use of corpora in natural language generation. In Section 3, we next describe the generation task on which the most corpus-driven work has been done: generating non-verbal behaviour for an embodied conversational agent. After that, we summarise in Section 4 several issues that must be taken into consideration when designing a corpus-based generation system, using specific examples from the conversational-agent task. Finally, in Section 5, we give some conclusions and recommendations.

2 State of the Art

In order to fully appreciate the specific issues that arise in multimodal corpus-based generation, it is necessary to understand the related work in multimodal corpora and in corpus-based text generation. This section summarises the current state of the art in these two research areas.

2.1 Multimodal Corpora

A multimodal corpus is a recorded and annotated collection of communication modalities such as speech, gaze, hand gesture, body language, generally based on recorded human behaviour.[1]

[1] Although Chafai et al. (2006) used a corpus based on Tex Avery cartoons.

Recently, researchers in this area have increasingly been coming together to share raw and annotated data, as well as techniques and tools for annotation and analysis. At the most recent in a series of workshops on multimodal corpora (Martin et al., 2006), a number of papers were presented describing corpora and their applications in areas including meeting analysis, hand gestures, multimodality during conversation, and multimodal human-computer interaction.

The normal method for annotating a multimodal corpus is to annotate each of the individual communication modalities on its own layer, and to make explicit or implicit links between the layers. Standard tools for doing this type of annotation include NXT (Carletta et al., 2005), Anvil (Kipp, 2004), and ELAN (Hellweig and Van Uytvanck, 2006). The types of data that are annotated depend both on the corpus and the intended applications, and may range from low-level time-stamped motions to high-level discourse structures. For example, the raw data for the AMI meeting corpus (Carletta, 2006) consists of 100 hours of recorded multi-party meetings, including full video and audio recordings of all participants, with fully-transcribed and time-stamped speech. The data has been annotated on the following levels: dialogue acts, topic segmentation, abstractive and extractive summaries, named entities, individual actions and gestures, person location, focus of attention, emotional content, and argumentation structure. Many of these levels are linked directly to segments of the transcript, while others—such as gestures—are marked with starting and ending times.

At the moment, many multimodal corpora are built and used mainly for descriptive purposes such as analysis and summarisation. For example, the primary applications of the AMI meeting corpus include human-human communication modelling, multimedia indexing and retrieval, and meeting structure analysis and summarisation. Most papers in Martin et al. (2006) describe such applications; however, multimodal corpora have also been used for generating output, particularly for embodied conversational agents. For example, Kipp et al. (2006) use a corpus to generate gesturing behaviour. This work is discussed in more detail in Section 3.

2.2 Corpora in Natural Language Generation

In the then-current state of the art in natural language generation summarised by Reiter and Dale (2000), the primary purpose of a corpus was to serve as guidance for human developers of a generation system: the texts in a corpus were used as targets to help in specifying the rules or target outputs of the system, but were not themselves used directly in the process of creating or evaluating the output.

In recent years, the increasing availability of large textual corpora, both annotated and unannotated, has contributed to the explosive development of computational-linguistics techniques that make direct use of the data represented in a corpus. The areas where data-driven techniques have been successful include machine translation, part-of-speech tagging, parsing, chunking, and summarisation (Manning and Schütze, 1999).

Researchers in natural language generation have now also begun to adapt these data-driven techniques. Modern data-driven NLG systems make use of textual corpora in two ways. On the one hand, corpus data can act as a resource for decision-making at all levels of the generation process; on the other hand, the data can also be used to help evaluate the output of a generation system. The work presented at a recent workshop (Belz and Varges, 2005) includes generation systems that employ corpora in both of these roles.

Using corpus data directly in the generation process has several benefits. First, it provides a means for making decisions that are difficult to encode in rules, but that can easily be derived from data. The corpus can be used to control the entire generation process: Marciniak and Strube (2005), for example, used machine-learning classifiers trained on a corpus of route descriptions to make all of the decisions in generation. It is also possible to integrate corpus-based models into more traditional generation frameworks. Williams and Reiter (2005) used corpus data to create rules for content selection; at the other end of the generation pipeline, the OpenCCG surface realiser (White, 2005), for example, uses n-gram language models as a resource for making decisions such as adverb placement within a rule-based framework. Incorporating data-driven variation into the generation process can also produce output that is less repetitive and that is often preferred by human judges (e.g., Belz and Reiter, 2006; Foster and Oberlander, 2006).

In additionto being used in the generation process, corpus data can also be used to evaluate the output of a generation system, generally by measuring how close the generated output comes to the texts in the corpus. Note that there is a danger in using cross-validation alone to evaluate the output of a generation system. As pointed out above, human judges in several studies (Belz and Reiter, 2006; Foster and Oberlander, 2006) have been found to prefer output that includes data-driven variation; however, a pure cross-validation measure will penalise such outputs against those that do not diverge far, on average, from the contents of the corpus, giving a potentially false picture of the relative quality. However, cross-validation and other corpus-driven methods can still provide a useful and easily computed evaluation of output quality and system performance, and have been used to evaluate a number of systems. For example, White (2004) measured the accuracy and speed of the OpenCCG surface realiser through cross-validation against target texts; Marciniak and Strube (2005) also evaluated their realisation component through cross-validation; Wan et al. (2005) used cross-validation to measure the recall and precision of a stochastic summary-sentence generation system; while Karamanis and Mellish (2005) describe a number of corpus-based methods for evaluating information-ordering systems.

3 Generating Non-Verbal Behaviour for ECAs

For the rest of the paper, we will concentrate on the specific task of generating multimodal behaviour for embodied conversational agents, as that is the target for most current data-driven multimodal generation systems. To be sure, corpora have been used in other multimodal generation systems as a resource for developers, à la Reiter and Dale (2000)—Corio and Lapalme (1999) used a corpus of information graphics and their captions to help define the rules for their system, for example—but it does not appear that corpora have been used directly for any multimodal generation task other than embodied agents, so we will focus on that task here.

Embodied Conversational Agents (ECAs) are computer interfaces that are represented as human bodies, and that use their face and body in a human-like way in conversation with the user (Cassell et al., 2000). The main benefit of ECAs as a user-interface paradigm is that they allow users to interact with a computer in the most natural possible setting: face-to-face conversation. However, to take full advantage of this benefit, the conversational agent must produce high-quality output, both verbal and non-verbal. A number of existing systems have based the choice of non-verbal behaviours for an ECA on the behaviours of humans in conversational situations; the implementations vary as to how directly they use the human data.

In some systems, motion specications for the agent are created from scratch, using rules derived from studying human behaviour; this is similar to the classical Reiter and Dale view of the role of corpora in text generation. For the REA agent (Cassell et al., 2001a), for example, gesturing behaviour was selected to perform particular communicative functions, using rules based on studies of typical North American non-verbal displays. Similarly, the performative facial displays for the Greta agent (de Carolis et al., 2002) were selected using hand-crafted rules to map from affective states to facial motions.

In contrast, other ECA implementations have selected non-verbal behaviour based directly on motion-capture recordings of humans. Stone et al. (2004), for example, recorded an actor performing scripted output in the domain of the target system. They then segmented the recordings into coherent phrases and annotated them with the relevant semantic and pragmatic information, and combined the segments at run-time to produce complete performance specications that were then played back on the agent. Cunningham et al. (2005) and Shimodaira et al. (2005) used similar techniques to base the appearance and motions of their talking heads directly on recordings of human faces. This technique can produce extremely naturalistic and individual output; however, the technical requirements for doing the motion capture are high, and the procedure is quite invasive for the subject.

A middle ground between the above two implementation strategies is to use a purely synthetic agent—one whose behaviour is controlled by high-level instructions, rather than based directly on human motions—but to create the instructions for that agent using the data from an annotated corpus of human behaviour. Like a motion-capture implementation, this technique can also produce increased naturalism in the output over a purely rule-based system, and also allows choices to be based on the behaviour of a single individual if necessary. However, annotating a video corpus can be less technically demanding than capturing and directly re-using real motions, especially when the corpus and the number of features under consideration are small. This approach has been taken, for example, by Cassell et al. (2001b) to choose posture shifts, by Foster and Oberlander (2006) to select facial displays, and by Kipp et al. (2006) to select hand gestures.

4 Designing a Multimodal Corpus for Generation

As described in Section 2, both multimodal corpora and corpus-based generation are currently active and productive areas of research. However, bringing together these two areas for corpus-based multimodal generation raises several issues that do not arise, or that do not have the same impact, in the two individual research areas: corpus-based techniques for text generation do not necessarily apply directly to the multimodal case, while general-purpose multimodal corpora are not always suitable for use in a generation system. The considerations when designing a corpus-based multimodal generation system include the following:

1. The contextual information necessary for making generation decisions must be represented in the corpus.
2. The granularity of the annotation and of the cross-modal links must be appropriate to the generation task.
3. The generation system must be able to reproduce the corpus data in an appropriate way.

In the remainder of this section, we will discuss each of these issues in more detail.

4.1 Representing Contextual Information

In many cases, multimodal corpora are created based on naturally-occurring human behaviour; that is, the subjects being recorded are free to speak and act as they wish, and the annotators then analyse the behaviour based only on the recordings. The corpus resulting from such a recording cannot contain any more information than what is available from observing the behaviour, and—possibly—from annotators applying their own judgement to add extra information (such as the dialogue-act and topic-structure annotations on the AMI corpus).

For some generation contexts, this sort of surface-level annotation of context is sufficient; for example, for an ECA whose motion is selected entirely based on the features of the speech signal, such as that of Shimodaira et al. (2005), no deeper representation of the context is needed. However, in many cases, a generation system has available a much richer notion of context as it is planning its output. For example, Greta (de Carolis et al., 2002) represents the target information structure and affective content of its utterances, while the input to the talking head of Foster and Oberlander (2006) includes the intended prosodic, dialogue-history, and user-model contexts. All of this information can be useful in choosing the desired multimodal output behaviour; however, unless it is represented in the corpus, none of it can be used by the generation system.

The required contextual information can be included in the corpus in two ways. Either it can be manually added after the fact by annotators, or the corpus can be created in such a way that the required information is already present before the annotation. The latter can be achieved by using corpora based on scripted output in the domain of the eventual target system; if the human being recorded is following a known script, then all of the relevant contextual information can easily be added to the corpus at construction time. This approach was taken by Stone et al. (2004) and Foster and Oberlander (2006). It has the advantage that no additional manual effort is required; however, it also has the disadvantage that the corpus must be created specifically for the target application, which rules out using existing annotated corpora.

4.2 Representing Cross-Modal Links

In many multimodal corpora, each separate modality is represented on its own timeline, with the only links between modalities those that are implicitly represented by the timestamps. For example, in the AMI corpus, there are many levels of links corresponding to different aspects of the spoken signal; however, the gesturing behaviour is represented on its own timeline with its own start and stop times. This type of representation is adequate if the goal is to extract events or to analyse human behaviour. However, if the goal is to generate novel output based on the corpus, more explicit links between the modalities are useful, as the temporal structure may not coincide with the underlying generation process.[2]

One important decision is the level at which cross-modal links are represented—that is, the size of the segment on each channel that can be associated with segments on other channels. For example, when associating multimodal behaviours with speech, motions may be associated with phonemes or syllables, with single words, with syntactic constituents, or with arbitrary sequences of words. Which of these is chosen depends on the level at which the generation system will later be selecting these motions; if the assumptions are later changed, it may prove costly. For example, the original talking-head implementation described by Foster and Oberlander (2006) selected facial displays based on individual words in the output, and the corpus was annotated accordingly. However, that assumption proved to be unrealistic: the majority of displays did *not* in fact coincide with single words. In order to produce more realistic motions, the entire corpus had to be re-processed using a revised scheme that associated displays with word sequences, and the generation system was updated to use that updated corpus.

As well as the level of representation, the criteria for making a link must be established: is the choice based strictly on temporal or spatial coincidence, or is semantic information also used? The former is easier to annotate, and may even be automatically derived from an existing annotated corpus, but may not generalise as readily to new outputs; the latter requires a more involved annotation process that makes more demands on the annotators for careful judgement calls. For example, Kipp et al. (2006) chose to record temporal co-occurrence and lexical affiliation as separate attributes when annotating hand gestures for generation; temporal co-occurrence was derived largely automatically from the video, but annotating the lexical links relied on "gesture literature and sometimes intuition".

4.3 Reproducing Corpus Data

A unimodal corpus can be used for generation with a minimum of processing effort; in most cases, the data in the corpus can be simply be directly combined to produce the output. For example, when using a textual corpus to help make decisions in surface realisation, n-gram models can be built from the words in the corpus and used to guide the system towards high-scoring realisations, as was done by Langkilde-Geary (2002) and White (2005). Similarly, in speech synthesis, the technique of unit selection (Hunt and Black, 1996) involves segmenting recorded speech into diphones (phoneme-to-phoneme transitions) and then using a Viterbi-style algorithm to construct a sequence of diphones to synthesise a given string of words. The corpus data must be annotated with the contextual information necessary to select the right content in a given context; however, there is no need to do any processing on the actual data to use it for generation.

For multimodal generation, in contrast, it is generally not the case that corpus data can be directly combined to produce output in the way that diphones can be concatenated for speech synthesis, or words for text generation. In most cases, a multimodal generation system creates entirely synthetic output by specifying commands for each of the relevant output channels, rather than combining existing pieces of output directly. Even in cases where motion-capture data is used directly (e.g., Stone et al., 2004; Cunningham et al., 2005), the recorded motions must still be mapped to animation commands and synchronised with the speech. When the generated output is specified at a higher level, then more complex mappings must be made.

For example, Kipp et al. (2006) use an annotation scheme for hand gestures that makes the conscious decision not to represent every single feature of the motion, but rather to capture

[2] Kipp et al. (2006) "found that the claim that gesture stroke and lexical affiliate always co-occur is often wrong."

the essentials, in some cases using gesture "lexemes" to abstract over the data. To recreate a gesture schedule annotated using this scheme, the motion specifications are translated into specific commands for the animation engine. Foster and Oberlander (2006) use a similar mapping for their facial displays: based on the speech, a set of high-level displays are selected, and are then converted to motion specifications in the language of the talking head.

It is important that the final mapping between annotated events and output commands is sufficiently close that the corpus data is actually relevant to the generation task. If not, the resulting output may not be appreciated by the subjects. For example, Foster (2004) attempted to use an annotated corpus of humans making hand gestures to specify the motion of an on-screen pointer, with rather disappointing results on the human evaluation.

5 Conclusions

Both multimodal corpora and corpus-based generation are active areas of research at the moment. Large-scale multimodal corpus resources such as the AMI corpus are being created and made freely available, and it would be a positive development if the data-driven techniques being developed for text generation could be directly applied to multimodal generation and could make use of such available resources.

Unfortunately, in many cases, the additional requirements of a multimodal generation system mean that it makes more sense in practice to collect and annotate a special-purpose corpus for the specific generation task, instead of using existing corpora. An application-specific corpus has the advantage that it can be created entirely from in-domain recordings, possibly even based on scripts to ensure that the necessary contextual information is readily available. Also, care can be taken that the data in the corpus is represented at the correct level for use in the generation system and that the output generator is able to make coherent output by using the data in the corpus.

However, it seems a shame to disregard entirely the corpora that are now being created, and it may often be possible to adapt such a corpus for use in a particular generation task. If an existing corpus is to be used for generation, it will likely be necessary to do some additional annotation to incorporate the necessary contextual and cross-modal information, and to take care in the implementation that the corpus data can be easily—and sensibly—reproduced. However, in some cases, this extra effort may be justified if it allows the generation system to take advantage of the increasing range of multimodal data to improve the generation process or to produce higher-quality output.

References

Belz, A. and Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).*

Belz, A. and Varges, S., editors (2005). *Corpus Linguistics 2005 Workshop on Using Corpora for Natural Language Generation.*

Carletta, J. (2006). Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5. Corpus available through `http://corpus.amiproject.org/`.

Carletta, J., Evert, S., Heid, U., and Kilgour, J. (2005). The NITE XML toolkit: Data model and query. *Language Resources and Evaluation Journal*, 39(4):313–334.

Cassell, J., Bickmore, T., Vilhjálmsson, H., and Yan, H. (2001a). More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1–2):55–64.

Cassell, J., Nakano, Y., Bickmore, T. W., Sidner, C. L., and Rich, C. (2001b). Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001).*

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (2000). *Embodied Conversational Agents.* MIT Press.

Chafai, N. E., Pelachaud, C., and Pelé, D. (2006). Analysis of gesture expressivity modulations from cartoons animations. In Martin et al. (2006).

Corio, M. and Lapalme, G. (1999). Generation of texts for information graphics. In *Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG'99)*, pages 49–58.

Cunningham, D. W., Kleiner, M., Wallraven, C., and Bülthoff, H. H. (2005). Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception (TAP)*, 2(3):251–269.

de Carolis, B., Carofiglio, V., and Pelachaud, C. (2002). From discourse plans to believable behavior generation. In *Proceedings of the 2nd International Conference on Natural Language Generation (INLG 2002).*

Foster, M. E. (2004). Corpus-based planning of deictic gestures in COMIC. In *Proceedings of the INLG 2004 Student Session*, Brockenhurst, England.

Foster, M. E. and Oberlander, J. (2006). Data-driven generation of emphatic facial displays. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).*

Hellweig, B. and Van Uytvanck, D. (2006). EUDICO linguistic annotator (ELAN) version 2.6: Manual. `http://www.mpi.nl/tools/`.

Hunt, A. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP96)*, volume 1.

Karamanis, N. and Mellish, C. (2005). A review of recent corpus-based methods for evaluating information ordering in text production. In Belz and Varges (2005).

Kipp, M. (2004). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation.* Dissertation.com.

Kipp, M., Neff, M., and Albrecht, I. (2006). An annotation scheme for conversational gestures: How to economically capture timing and form. In Martin et al. (2006).

Langkilde-Geary, I. (2002). An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 2nd International Language Generation Conference (INLG 2002).*

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press.

Marciniak, T. and Strube, M. (2005). Using an annotated corpus as a knowledge source for language generation. In Belz and Varges (2005).

Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., and Pianesi, F., editors (2006). *LREC 2006 Workshop on Multimodal Corpora: From Multimodal Behaviour Theories to Usable Models.*

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge University Press.

Shimodaira, H., Uematsu, K., Kawamoto, S., Hofer, G., and Nakai, M. (2005). Analysis and synthesis of head motion for lifelike conversational agents. In *Proceedings of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2005).*

Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Lees, A., Stere, A., and Bregler, C. (2004). Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513.

Wan, S., Dale, R., Dras, M., and Paris, C. (2005). Statistically generated summary sentences: A preliminary evaluation using a dependency relation precision metric. In Belz and Varges (2005).

White, M. (2004). Reining in CCG chart realization. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG 2004)*.

White, M. (2005). Designing an extensible API for integrating language modeling and realization. In *Proceedings of the ACL 2005 Workshop on Software*.

Williams, S. and Reiter, E. (2005). Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application. In Belz and Varges (2005).

Towards a Cognitive Model of Multimodal Output for Language Production*

Markus Guhe
Human Communication Research Centre, University of Edinburgh
Adam Ferguson Building, 40 George Square
Edinburgh, UK EH8 9LL
m.guhe@ed.ac.uk

Abstract

In this paper I discuss some points relevant for developing a computational cognitive model of the conceptualiser, the first component in the human language production system according to Levelt (1989). I describe extensions of the *incremental conceptualiser* (INC) with the aim of creating a cognitively plausible multimodal dialogue model. I point out a way to represent multimodal knowledge on the conceptual level, and I discuss at which point towards the production of preverbal messages (semantic structures) modal fission, i.e. the split into different modalities, occurs. I argue for a late modal fission; that is, fission does not occur during Levelt's macroplanning but during microplanning.

Keywords: language production, multimodality, cognitive modelling, incremental conceptualisation

1 Introduction

Humans communicate in different modalities. The most notable and important one is without doubt language, but human communication also makes use of other modalities like gaze, intonation, gesture or body posture. Although communication can succeed with language alone, for example, in telephone conversations, communication makes use of other modalities in the most natural setting: face-to-face dialogue, cf. Clark (1996). Much progress has been made in recent years in creating artificial, computer-animated multimodal dialogue agents that act in as natural a way as possible. These systems are mostly in the tradition of Artificial Intelligence, which means they are created with the goal of being systems that exhibit intelligent, human-like behaviour. While these systems rapidly approach an everyday-use quality, one issue is often neglected: how does the human system function? This question is relevant for two main reasons. Firstly, there is the primeval scientific urge to understand humans, and because language is such a central property of what makes humans human, any improvement in understanding the functioning of language will further the understanding of what it means to be human. Secondly, the gold standard, against which artificial conversational agents have to be evaluated is the way humans communicate, and a better understanding of the cognitive processes underlying language offers new insights into how systems can act more naturally. One aspect of this approach is that mechanisms operating in humans can be implemented in artificial systems, which may improve these systems; something, which can be called 'bionics for AI'.

Notable exceptions to the predominance of non-cognitive approaches to multimodal communication are the models proposed by Krauss et al. (2000) and de Ruiter (2000). (However, neither model is implemented as a computational system.) Both are models of how speech and gesture are

*The research reported here was supported by a grant from the National Science Foundation (IIS-0416128). Thanks to Mark Steedman, Ellen Bard, Max Louwerse and the anonymous reviewers for comments on earlier versions.

generated in a coordinated fashion. The models are quite different (see, for example, Feyereisen 2006 for a comparison). The main difference is that the Krauss et al. model assumes two independent processes – one for gesture planning, one for linguistic planning – that operate directly on working memory and that are only coordinated at the phonological planning stage. De Ruiter on the other hand extends the conceptualiser (the first component in the language production model by Levelt 1989) to also generate gesture *sketches* in direct conjunction with the propositional preverbal messages (semantic structures) that encode the linguistic content to be communicated.

In this paper I take a cognitive modelling approach to multimodal output generation. That is, the aim of this line of research is to build a computational model of how humans produce multimodal contributions to a dialogue. More precisely, I describe a way to represent knowledge in different modalities on the conceptual level, and I narrow down the point in the conceptualisation process at which modal fission, that is, the split into different modalities, takes place. I argue for a late modal fission that is integrated with the generation of preverbal messages.

2 Cognitive Modelling

Building a computational cognitive model of multimodal communication relies on the information-processing approach to cognition. This paradigm says that perception and cognition consists of processes of information-processing, which has the consequence that these processes can be modelled and simulated by algorithms and be executed on computers (Newell, 1990; Simon, 1996). Whereas this is similar to the main goal of AI to build systems capable of exhibiting intelligent behaviour, in the field of cognitive modelling there is an additional constraint: the algorithms modelling perception and cognition not only should produce adequate output for corresponding input, but the algorithms also should operate in the same way as human cognitive processes. In his discussion of cognitive algorithms Steedman (1998) points out two properties of cognitive algorithms that are particularly relevant for the approach I take in the *incremental conceptualiser* (INC, see Section 3). The first is that cognitive algorithms must be simple algorithms. That is, their complexity must be low enough for the computation to finish in a very limited time-window (in 'real time'). Secondly, the algorithm must be capable of dealing with realistic average case scenarios, i.e. it must have a good average case performance.[1] For example, it is rather obvious that NP-complete algorithms cannot be considered cognitively adequate: because of their complexity, they do not produce results in an adequate time (except for the tiniest of cases). These are strong arguments for the resource-saving incremental processing mechanism used in INC.

Cognitive modelling assumes a triangle of theory-formation, model-building, and empirical validation. First a theory about an aspect of human cognition is formed, which is then cast into a computational model, i.e. the model is being implemented. Simulations carried out with the implementation are compared to experimental data, and the discrepancies between model and data lead to a refinement of the theory, which leads to improvements of the model. In this paper I present preliminary work aimed at building a model of multimodal output generation that will then be tested against human data. The data is collected in a modified Map Task (Anderson et al., 1991; Guhe et al., 2006) and is currently evaluated.

It is important to note that the aim of cognitive modelling is to *model* perceptual and cognitive processes, not to recreate them. Due to the large number of cognitive processes operating at any given time in a cogniser and due to the many details that would have to be implemented on the low levels of perception and cognition it is difficult, if not impossible to build models that fit the human behavioural data perfectly on every level of abstraction. So deciding which aspects to leave out of the model is almost as important as finding the best fit to the aspects that are being modelled.

Within cognitive modelling there is the tradition of developing unified theories of cognition (Newell, 1990; Anderson and Lebiere, 1998). Although the search for the 'atomic components of thought' (Anderson and Lebiere, 1998) surely is a desirable undertaking, for the question at hand, namely, to build a model of multimodal output generation, the existing theories and architectures

[1] Although worst-case scenarios are not particularly relevant, breakdowns of the cognitive function can reveal aspects of the algorithm's functioning.

are not developed far enough. The main problems are that the existing unified theories of cognition do not have a particularly detailed account of language. One of very few exception is the language processing model in Soar by Lewis (see Lewis 1993; Lewis and Vasishth 2005). However, these exceptions do not address issues of language generation, and they are too general to be used for the purposes I am addressing here.

3 Conceptualisation

3.1 Levelt's conceptualiser

The conceptualiser is the first of the three components in Levelt's (1989) model of language production. It performs different sub-tasks to generate semantic structures (termed *preverbal messages* by Levelt) for a communicative intention. Two major tasks of the conceptualiser are *macroplanning* and *microplanning*. Macroplanning comprises the steps a speaker performs to decide upon the content that needs to be communicated in order for the dialogue partner(s) to recognise the communicative intention. Microplanning consists of the steps necessary to turn the selected content into the appropriate format so that the formulator – the component receiving the preverbal messages – can generate spoken (or written) language expressing the communicative intention.

In Section 4 I argue that message generation proceeds through most of the conceptualiser without fission into modalities. Only when the output for the content determined by macroplanning is generated during the microplanning stage of conceptualisation, the (partial) fission into different modalities occurs. Whereas de Ruiter (2000) suggests that fission occurs at the level of macroplanning, I propose that at the very least this is not true for all modalities. (I discuss intonation, gaze, eyebrow raises and pointing gestures.) But before making this argument, I give a more detailed account of conceptualisation in this section by describing the *incremental conceptualiser* (INC, Guhe 2007a; Guhe and Habel 2001), which is an implemented computational model of the conceptualiser, and describe how it can be used to process multimodal knowledge. INC can serve as framework to work out the details of how and where output for modalities other than language is generated. By doing this it will be extended into a multimodal model. The main argument I make in this paper is that in INC the question where modal fission occurs is whether the algorithms of the selection process (macroplanning) or the ones of the preverbal-message-generation process (microplanning) generate output in the non-linguistic modalities.

3.2 The Incremental Conceptualiser

The incremental conceptualiser (INC; Guhe 2007a; cf. Figure 1) is a computational cognitive model of the first component of Levelt's (1989) model of language production. It is located between a preprocessing unit (PPU), which computes simple concepts from perceptual input, and the formulator, which does the linguistic encoding. INC has four main processes that correspond to Levelt's four main stages of conceptualisation:

1. Construction (Levelt's bookkeeping) receives input from the PPU and incrementally builds up the current conceptual representation (CCR)[2]. Construction uses the concept matcher, which accesses conceptual patterns in the concept storage to build up hierarchical conceptual structures.

2. Selection (macroplanning 1) selects the situations from the CCR that will be verbalised in order to realise the overall communicative intention. By doing so, it generates 'sub-intentions' (Guhe, 2007a).

3. Linearisation (macroplanning 2) brings the selected situations into an appropriate order.

4. PVM-generation (preverbal message generation, Levelt's microplanning) incrementally generates preverbal messages for each selected situation by deciding which concepts are verbalised in which way.

[2] It is a *current* conceptual representation, because the representation changes over time, i.e. all processing takes place on the current state of the representation.

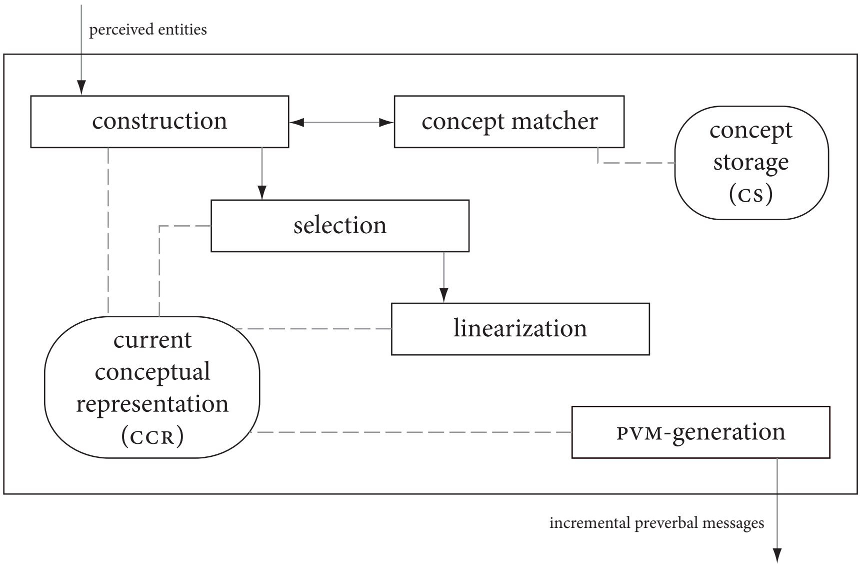


Figure 1: The incremental conceptualiser

INC's processes work incrementally, that is, they work in a piecemeal fashion. The processes are arranged in a cascade, which means that they run in parallel but that the output of one process is the input to its successor in the cascade. (This is what Reiter (1994) calls a pipeline with the difference that the processes work in parallel.) For example, a situation can only be selected for verbalisation after it has been inserted into the CCR by construction. Although the incremental mode of operation in this cascaded architecture is one of the central properties that make INC a cognitive model, it is not the major concern of this paper. Guhe (2007a) contains a full discussion of INC's incrementality.

The difference between selection and PVM-generation is an important one: selection decides *that* a situation is verbalised, whereas PVM-generation decides *how* the situation is described.[3] Verbalising a situation requires further concepts, which are chosen by PVM-generation. This distinction makes it possible to ask the main question of this paper more precisely: does modal fission take place during the selection/linearisation phase (macroplanning) or during PVM-generation (microplanning)?

For the future development of INC this question is important, not only because it will determine in which component the fission algorithm(s) will be located but also because it has important consequences for the kind of representations used as well as for the kind of algorithms needed.

3.3 A Multimodal CCR

The only requirements that INC has about the structure of its main representation, the CCR, are that it must contain information about situations and that the knowledge is represented in an 'object oriented' way; that is, the conceptual knowledge must be organised around concepts. Whether the information is purely propositional or whether there is knowledge available in other modalities as well (or whether the representation is amodal instead of multimodal) plays only a secondary role.

INC can, therefore, be extended straightforwardly to operate with a conceptual knowledge representation that includes knowledge about the represented concepts in other modalities. Most researchers agree that knowledge represented in working memory and in long-term memory is multimodal and in addition to propositional includes imagistic and kinaesthetic knowledge (Kosslyn,

[3] This division of labour is similar to the one between *what-to-say* and *how-to-say* (de Smedt et al., 1996), but I only refer to a division of labour within the conceptualiser.

1995; Levelt, 1989). It also seems to be clear that the preverbal message must be a purely propositional representation to be turned into linguistic output. Therefore, one task the conceptualiser has to perform is to translate the multimodal knowledge in the conceptual representation into propositional structures for the preverbal message.

The CCR is realised as a referential net (Habel, 1986). All knowledge is organised around interlinked referential objects (refOs), which in the representation are proxies for entities. In the CCR each concept is represented by a refO. Attributes and designations associated with the refO represent the knowledge about the entity. Attributes represent conceptual knowledge, which is, among other things, used for performing inferences and deductions; designations represent meaning-related knowledge, that is, knowledge about how to talk about the entity.

Spatial knowledge can be represented by designations and attributes, and Habel (1987) describes how referential nets can be extended to represent pictorial or imagistic knowledge. (See Landau and Jackendoff (1993) on the distinction between spatial and imagistic knowledge.) This is done by extending the basic referential nets consisting of refOs, attributes and designations by depictions. Although it has not been done yet, knowledge in other modalities could be represented in an analogous way, for example, the building blocks for de Ruiter's (2000) sketches. The simplest of these building blocks would be iconic gestures, where a specific gesture can be used to refer to a particular concept.

Introducing these additional representational means into the originally purely propositional referential nets formalism touches on a fundamental problem of formal representations: to describe a cognitive representation in a non-propositional modality, one has to use some form of representational language. Simply using a picture to represent imagistic knowledge is not very helpful, because a picture does not specify how it can be used in algorithms. Algorithms operating on the picture must be specified in some (formal) language, which means by propositional means. However, the knowledge that is described by these propositional means need not itself be propositional. Thus, this extended form of referential nets can be used for a multimodal version of INC, where the CCR contains knowledge in different modalities, but these representations are suitable to be processed by symbolic, propositional algorithms.

4 Locating Fission for Four Modalities

In this section I discuss the place in the conceptualiser where modal fission must take place for four modalities: intonation, gaze, eyebrow raises and pointing gestures. I argue that – for these modalities at least – fission must occur late in conceptualisation, because the decisions which modalities to use must be made in conjunction with generating the semantic representation (or, the information structure) that underlies utterances.

4.1 Intonation

In the framework of his Combinatory Categorial Grammar Steedman (2000, 2006) argues that theme and rheme of the information structure underlying an utterance determine the production of pitch accents in English.

> A theme is a part of the meaning of an utterance that the speaker claims some participant in the conversation supposes (or fails to suppose) already to be common ground; ... A rheme is a part of the meaning of an utterance that the speaker claims some participant in the conversation makes (or fails to make) common ground. (Steedman, 2006)

Thus, a theme is the part of the meaning of an utterance that the speaker supposes to be already part of the common ground (roughly, the 'Question Under Discussion') and a rheme is the part of the meaning of an utterance that the speaker makes common ground through this communicative act (roughly, the answer to the question). The 'common ground consists in the set of propositions that a conversational participant supposes to be mutually agreed to for the purposes of the conversation.' This is 'not necessarily the same as the set of propositions that all participants actually believe' (Steedman, 2006). Following Rooth (1992), Steedman demonstrates that themes and rhemes depend on the alternatives a speaker has at the point in the dialogue where he or she makes the contribution to the dialogue.

Similar to the case that pitch accents are a result of the theme–rheme structure, the boundary tones required to determine the full intonational contour of the utterance depend on polarity; that is, they depend on whether the speaker or hearer is or is not supposing the theme/rheme 'to be or to be made common ground – that is, whether it is contentious or uncontentious' (Steedman, 2006). Preverbal messages that are specified in this way result in utterances like

```
There is a RED church and a BLACK church.
           L+H*    LH%          H*          LL%
```

Where 'red' is the focused item that is the theme of the utterance and 'black' the rheme. (This utterance makes a contrast to the statement by the dialogue partner: 'There is a red church and a white church.')

In Guhe (2006, 2007b) I show how INC can be extended to generate preverbal messages that have an appropriate marking of theme and rheme by a combination of Rooth's alternatives semantics and the 'magic circle'. The magic circle is a useful notion in explaining the Map Task that, given a location on a map, establishes roughly the area of the map under discussion and that contains the prime places that the dialogue partners will talk about next. While it is generally difficult to determine alternatives sets (Cohen, 1999), for the Map Task they are closely linked to the magic circle. In INC the alternatives sets are computed by the selection process, because the alternatives sets contain the propositions that a speaker could have chosen as the next contribution to the dialogue (Guhe, 2007b, 2006). Theme and rheme are marked in the preverbal message by adding attributes to the output refOs containing the theme/rheme information.

Furthermore, what the speaker assumes to be common ground is just a sub-structure of the CCR and, analogously, what the speaker assumes the hearer assumes to be common ground. By computing whether the speaker assumes that the planned utterance will (not) become part of the speaker's or the hearer's common ground and by adding this information to the preverbal message as well, the boundary tones are determined.

An empirical validation of these claims, that is, whether the pitch accents and boundary tones do actually have these functions, is currently underway in new experiments using a new version of the Map Task.

Concluding, intonation can be computed solely on the grounds of the information structure of a preverbal message (what Steedman would call a semantic representation). The consequence for the question of where modal fission is located in the conceptualiser is that the intonational modality is simply a consequence of the allocation of theme and rheme in the preverbal message and a consequence of the status of the associated information in common ground (a consequence of polarity). Thus, the relevant information is generated as part of microplanning.

4.2 Gaze

Cassell et al. (1999) present data of a study that correlates gaze with the theme–rheme structure of utterances. They find the following:

> Specifically, the beginning of themes are frequently accompanied by a look-away from the hearer, and the beginning of rhemes are frequently accompanied by a look-toward the hearer. When these categories are co-temporaneous with turn-construction, then they are strongly—in fact, absolutely—predictive of gaze behavior. (Cassell et al., 1999)

With the marking of theme and rheme already in place, the preverbal message now just needs to indicate whether it is the beginning of the speaker's turn.

This should not be taken to mean that turn-construction and the theme–rheme structure are the only factors in determining gaze behaviour, in particular if the two factors do not coincide. Additionally, gaze is not at the discretion of the language production system but is in determined by a multitude of factors. For example, an unexpected event can draw the attention of a speaker, which typically has the consequence of the speaker shifting his or her visual attention in the direction of the event.

Despite these caveats, the default gaze behaviour of a speaker is determined by the propositional content of the preverbal message.

4.3 Eyebrow raises

Eyebrow raises have often been associated with emphasis, contrast or pitch accents. Although experimental evidence for this is still sketchy and controversial, there is some evidence that eyebrow raises are correlated with the information structure of an utterance. Flecha-García (2006a,b) shows that eyebrow raises and pitch accents are generated in a coordinated fashion, although she finds no evidence that they actually mark contrast, as is often assumed. She concludes that the conversational functions of eyebrow raises are 'to signal the beginning of high-level discourse segments and to emphasise information in the utterances with the most important role in the dialogue.' (Flecha-García, 2006a, p. 1315)

Even if that would be all there is to eyebrow raises it points towards a system in which eyebrow raises depend on particulars of the information structure, not the utterance as a whole. So, this, again, favours micro- over macroplanning as location for modal fission.

4.4 Pointing gestures

Pointing gestures are deictic gestures where a speaker points to an object in the world in order to refer to it, typically by using his or her index finger. (A precise definition of these gestures is difficult and not urgent for the purpose at hand.) Van der Sluis (2005, see also van der Sluis and Krahmer 2001) presents a model of generating multimodal referring expressions that consist of a pointing gesture accompanied by a verbal referring expression. She demonstrates that the precision of the pointing gesture depends on how many distractors (other possible entities the referring expression is identifying) are not ruled out by the propositional content of the expression (the verbal referring expression). In other words, the more precise the pointing gesture, the less verbal output must be generated in order to uniquely refer to an object in the world and vice versa. Although van der Sluis's model is not a cognitive model as such, the behaviour of the algorithms is tested against behaviour observed in accompanying experiments. This means, the very least this model does is to give a good indication of what the cognitive algorithms look like. Most important in the context of this paper is the interdependence of the precision of the pointing gesture and the information that is mentioned in the referring expression. This means that both have to be generated by the same algorithm, and from this follows that this is the location where modal fission occurs.

A point to note here is that pointing gestures will most likely not be generated from preverbal messages, i.e. the instruction of where to point and how precise the gesture should be is not specified in the preverbal message, because gesturing requires imagistic and/or spatial representations instead of propositional ones. Thus, for pointing gestures the fission is part of microplanning and a representation akin to de Ruiter's (2000) sketch is generated that is sent to a gesture planner in synchrony with the preverbal message that is being sent to the formulator.

The conclusion is once again that the planning of the pointing gesture is a part of microplanning.

5 Conclusion

In this paper, I discussed some points relevant to building computational cognitive model of a multimodal conceptualiser. After describing a way to represent multimodal conceptual knowledge, I looked at some evidence concerning the question of where modal fission is located in the human language production system, more specifically: where in the conceptualiser. This mini-survey of available evidence is certainly far from being complete, but it is striking that in all four cases modal fission seems to take place during Levelt's (1989) microplanning, which stands in contrast to de Ruiter's (2000) proposal, who locates it as part of macroplanning; that is, earlier in the conceptualisation process. For the model INC this means that modal fission is a sub-task of the preverbal-message-generation process, not its selection process. (It is probably more appropriate to say that preverbal-message generation is a sub-task of an multimodal output generation process and that modal fission is part of the functioning of this general process.)

This should not be taken to mean that all modal fission must occur at this level. However, behaviour other than language that is used for communication is planned in conjunction with generating the propositional content of language. Behaviour in the modalities intonation, gaze and eyebrow raises depends on properties of the preverbal message's information structure. For pointing gestures possible trade-offs between gesture and speech can only be planned by the same algorithm; otherwise they would be simply random.

It should be noted that using a particular modality can be intended although it is not part of macroplanning. This means, for example that the results of Melinger and Levelt (2004), who argue that gestures are intended actions, do not interfere with planning of gestures being a part of microplanning. A speaker can also intend to use a particular word or decide to use a passive sentence, which extends the reach of intentionality even further down the language production system. (But it becomes more difficult for the speaker to intend something the further down the system a particular function is located, because of the stronger informational encapsulation of these components.) Macroplanning is the task that determines what I called 'communicative sub-intentions' in Guhe (2007a) – determining what to communicate. Microplanning is the step to determine the means with which these sub-intentions are realised, but this does not mean this level is non-intentional.

Finally, multimodal behaviour is certainly not completely determined by conceptualisation. In particular, conceptualisation does not need to determine completely which modalities are used in which combinations. Some decisions must be made in the modality-specific components that use the output of the conceptualiser. For example, marking theme/rheme and whether a piece of the preverbal message makes knowledge common ground does not explicitly address the issue of intonation, but the component that is located later in the language generation system and that generates intonation uses this information for generating an appropriate output. It may even turn out that the independent, autonomous decisions of this kind are the mechanism that gives rise to redundant use of multiple modalities.

References

Anderson, A. H., Bader, M., Bard, E. G., Bolyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Anderson, J. R. and Lebiere, C. (1998). *The Atomic Components of Thought.* Lawrence Erlbaum, Mahwah, NJ.

Cassell, J., Torres, O. E., and Prevost, S. (1999). Turn taking vs. discourse structure: How best to model multimodal conversation. In Wilks, Y., editor, *Machine Conversations*, pages 143–154, The Hague. Kluwer.

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

Cohen, A. (1999). How are alternatives computed? *Journal of Semantics*, 16:43–65.

de Ruiter, J. P. (2000). The production of gesture and speech. In McNeill (2000), pages 284–311.

de Smedt, K., Horacek, H., and Zock, M. (1996). Some problems with current architectures in natural language generation. In Adorni, G. and Zock, M., editors, *Trends in Natural Language Generation: An Artificial Intelligence Perspective*, pages 17–46. Springer, New York.

Feyereisen, P. (2006). How could gesture facilitate lexical access? *Advances in Speech–Language Pathology*, 8(2):128–133.

Flecha-García, M. L. (2006a). Eyebrow raising, discourse structure, and utterance function in face-to-face dialogue. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1311–1316.

Flecha-García, M. L. (2006b). *Eyebrow raising in dialogue: discourse structure, utterance function, and pitch accents*. PhD thesis, Theoretical and Applied Linguistics, University of Edinburgh.

Guhe, M. (2006). Intonational information in preverbal messages. In preparation.

Guhe, M. (2007a). *Incremental conceptualization for language production*. Lawrence Erlbaum, Mahwah, NJ. In press.

Guhe, M. (2007b). Marking theme and rheme in preverbal messages. In *Proceedings of the Seventh International Workshop on Computational Semantics*.

Guhe, M. and Habel, C. (2001). The influence of resource parameters on incremental conceptualization. In Altmann, E. M., Cleeremans, A., Schunn, C. D., and Gray, W. D., editors, *Proceedings of the 2001 Fourth International Conference on Cognitive Modeling: July 26–28, 2001, George Mason University, Fairfax, VA*, pages 103–108, Mahwah, NJ. Lawrence Erlbaum.

Guhe, M., Steedman, M., Bard, E. G., and Louwerse, M. (2006). Prosodic marking of contrasts in information structure. In *Proceedings of BranDial 2006: The 10th Workshop on the Semantics and Pragmatics of Dialogue, University of Potsdam, Germany, September 11th-13th 2006*.

Habel, C. (1986). *Prinzipien der Referentialität: Untersuchungen zur propositionalen Repräsentation von Wissen*. Springer, Berlin, Heidelberg.

Habel, C. (1987). Cognitive linguistics: The processing of spatial concepts. In TA *Informations, Bulletin semestriel de l'*ATALA*, Association pour le traitement automatique du langage, 28*, pages 21–56.

Kosslyn, S. M. (1995). Mental imagery. In Kosslyn, S. M. and Osherson, D. N., editors, *Visual Cognition*, volume 2 of *Daniel N Osherson (general editor), An Invitation to Cognitive Science*, chapter 7, pages 267–296. MIT Press, Cambridge, MA, 2 edition.

Krauss, R. M., Chen, Y., and Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In McNeill (2000), pages 261–283.

Landau, B. and Jackendoff, R. (1993). 'what' and 'where' in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–265. With commentary and response.

Levelt, W. J. (1989). *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Lewis, R. L. (1993). *An Architecturally-based Theory of Human Sentence Comprehension*. PhD thesis, Carnegie Mellon University. Also available as Computer Science Tech Report CMU-CS-93-226.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

McNeill, D., editor (2000). *Language and Gesture*. Cambridge University Press, Cambridge, MA.

Melinger, A. and Levelt, W. J. (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4(2):119–141.

Newell, A. (1990). *Unified Theories of Cognition: The William James Lectures, 1987*. Harvard University Press, Cambridge, MA.

Reiter, E. (1994). Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation (*INLGW*-1994)*, pages 163–170, Kennebunkport, ME.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Simon, H. A. (1996). *The sciences of the artificial.* MIT Press, Cambridge, MA, 3rd edition.

Steedman, M. (1998). Cognitive algorithms: Questions of representation and computation in building a theory. In Scarborough, D. and Sternberg, S., editors, *Methods, Models, and Conceptual Issues*, volume 4 of *Daniel N Osherson (general editor), An Invitation to Cognitive Science*, chapter 5. MIT Press, Cambridge, MA, 2 edition.

Steedman, M. (2000). *The syntactic process.* MIT Press, Cambridge, MA.

Steedman, M. (2006). Syntax, semantics, and implicature in English intonation. Draft 2.0 of 25 September 2006.

van der Sluis, I. (2005). *Multimodal Reference: Studies in Automatic Generation of Multimodal Referring Expressions.* PhD thesis, Tilburg University, The Netherlands.

van der Sluis, I. and Krahmer, E. (2001). Generating referring expressions in a multimodal context. In Daelemans, W., Sima'an, K., Veenstra, J., and Zavrel, J., editors, *Computational Linguistics in the Netherlands 2000. Selected Papers from the Eleventh* CLIN *Meeting*, No 37 of Language and Computers: Studies in Practical Linguistics (edited by Jan Aarts and Willem Meijs), pages 158–176. Rodopi, Amsterdam, New York.

On Reciprocal Improvement in Multimodal Generation: Co-reference by Text and Information Graphics[*]

Christopher Habel and Cengiz Acartürk
Department of Informatics
University of Hamburg
D-22527 Hamburg, Germany
{habel / acarturk }@informatik.uni-hamburg.de

Abstract

In this paper, we argue that in the production of complex documents combining text and information graphics, improvement—seen as a specific revision process—has to be considered as an independent module in a multi-pass architecture. Inside improvement, graph comprehension and text comprehension modules interact in building up a common content representation based on a conceptual inventory specified via topological and geometrical concepts. During concurrent comprehension the reciprocal improvement module inspects possible gaps in co-reference and coherence and decides which gaps should be filled. We exemplify these tasks and processes with an analysis of an excerpt from a business-news article in the New York Times.

Keywords: Multimodal generation, graph comprehension, conceptual representations, spatial concepts, information graphics, business news

1 COMBINING TEXT AND INFORMATION GRAPHICS

Documents containing modalities such as figures (graphs, drawings, photographs), tables, equations etc. together with text are wide-spread in print media as well as in electronic media. The most frequently cited argument for combining modalities—we call it the *argument of division of labor between representational modalities*—can be exemplified by a characterization given in the *Publication Manual* of the American Psychological Association (APA):

> Tables are often preferred for the presentation of quantitative data in archival journals because they provide exact information; figures typically require the reader to estimate values. On the other hand, figures convey at quick glance an overall pattern of results. They are especially useful in describing an interaction—or lack thereof—and nonlinear relations. A well-prepared figure can also convey structural or pictorial concepts more efficiently than can text. (4th ed., 1994, p. 141)

Although scientific texts—in particular those in a specific field—are the main focus of interest in this characterization, it is also applicable to multimodal documents in non-scientific journals or newspapers etc., as well as to corresponding documents accessible via the Internet. Interpretation of graphics is crucial in many areas such as trends in economy, diagnosis and medical treatment, time-series analysis of experimental data, computer-based instructional material in early school education, among others. Graph

[*] The research reported in this paper has been partially supported by the DFG (German Science Foundation) in ITRG 1247 'Cross-modal Interaction in Natural and Artificial Cognitive Systems' (CINACS).

comprehension research literature emphasizes a broad range of factors playing important role in interpretation of graphical data. For example, Peebles and Cheng (2001) specifies information retrieval from graphics as the interaction between visual properties of graph, cognitive abilities of graph user and requirements of task. There are also studies suggesting methods and design guidelines for improvements to facilitate reader's interpretation of graphics and identification of trends (e.g. Kosslyn, 1994, 2006). Nevertheless, there are inconsistencies in interpretation of the same graphical data (e.g. trends in time-series graphics) even among expert scientists (DeProspero and Cohen, 1979). Such limitations, caused by the implicit nature of diagrammatical representations to present certain aspects in interpretation of quantitative data as well as perceptual difficulties during early stages of information extraction bring the need for multimodal documents including both graphics and text.

While diagrammatical representations are accepted to be computationally more efficient than sentential representations in specific tasks and domains (Larkin and Simon, 1987), in addition to the proposals that use of multiple modalities facilitates learning (Ainley et al. 2000, Winn 1991), the principal pros of combining information in different modalities are opposed to—possibly—additional cognitive efforts of producers and recipients in processing cross-modal relations (e.g. co-reference, coherence etc.) and in considering cross-modal dependencies. Thus, systematic investigations of such relations and dependencies are fundamental for any approach to comprehension or production of multimodal documents combining text and graphics. In the present paper, we exemplify the additional cognitive tasks with the case of *co-reference*, which was also in the focus of some NLG approaches to generation of multimodal output, such as WIP or AutoBrief (for an overview see André, 2000).

2 REFERENCE AND CO-REFERENCE IN MULTIMODAL DOCUMENTS

From a linguistic point of view, the basic type of *referring* is constituted by a *referential expression* that *refers* to an *entity* of the domain of discourse. *Co-reference*, the backbone of text coherence has to be established by speaker and hearer employing internal—conceptual—representations, which mediate between the language and the domain of discourse. In processing multimodal documents, additional types of reference and co-reference relations have to be distinguished. Foremost, there exist corresponding referential relations (reference links) between graphical entities and entities in the domain of discourse.[1] Figure 1 shows a hand-drawn sketch map: some of the *lines* refer to rivers, to roads or to parts of the costal line, i.e. they refer to entities in the geographical world, whereas other lines constitute a *rectangle*, i.e. member of another class of graphical entities, which refers to a region, namely Aberdeen University's 'Old Aberdeen Campus', or other regions, such as part of a harbour or the North Sea (see Figure 1). In other words, the systems of regularities for combining atomic graphic entities to complex, meaningful configurations behave similar to grammatical systems.

When the producer of this sketch map explains the environment using the map, for example by saying "The rail station is between the red lines left of the harbour", the recipient has to integrate referential links of different types. Firstly, there are links between linguistic expressions and geographical entities, e.g. a reference relation between "*rail station"* and a *building in the town of Aberdeen*. Secondly, there are referential links between graphical entities, e.g. *lines*, and geographical entities, e.g. *streets*, and thirdly there are links between linguistic expressions and graphical entities, e.g. between "*red lines*" and some *red lines on a sheet of paper.* The composition—in the mathematical sense of composition of maps or relations—of the language-to-graphics reference and the graphics-to-domain reference leads to a language-to-domain reference, which we call in the following *implicit reference.* In particular, the implicit reference link mentioned above connects the phrase "*between the red lines*" to a region in the town of Aberdeen, which was not explicitly mentioned in the text.

[1] In the present paper we focus exclusively on co-reference in text–graphics multimodality; other types of multimodal communication, e.g. the combination of language and gesture, show similar phenomena and problems (see Beun and Cremers, 2001). Graphical entities that hold meaning with respect to the domain, and thus can be seen as corresponding to words, phrases or sentences in language, is one the central research topics in the psychology of graph comprehension (cf. Kosslyn, 1989; Shah and Hoeffner, 2002).

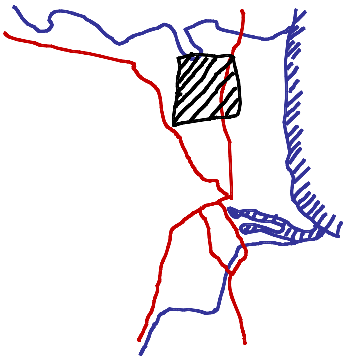


Figure 1: Sketch map of Aberdeen (pure depiction without textual elements).

The types of reference and co-reference relations we exemplified above with the class of (sketch) maps are central for analyzing the reference relations that are involved in the processing of text-depiction documents in general. (See Figure 2, which depicts the structure of different types of referential links and their composition.) Tappe and Habel (1998) describe that people producing verbal descriptions of the drawing of sketch maps employ two conceptual layers of representation: a layer corresponding to graphical entities and a layer corresponding to domain entities (entities of the real world referred to by text and sketch map). The empirical data of their experimental study supports the assumption that both layers are simultaneously accessible during speech production. Furthermore, different experimental settings of the verbalization task leads to different use of referential types, which results in different usage of words and phrases, correspondence to graphic entities vs. correspondence to real-world entities. In contrast to Tappe and Habel (1998), which focus on the relations between the external representations (language and graphics) and the layers of conceptual representations active in language processing, Tabachneck-Schijf, Leonardo and Simon's (1997) CaMeRa model of *Computation with multiple representations*, which proposes 'referent ties' as a major means to link different layers of internal representations, namely between pictorial and verbal short-term memory, focuses on the connections between the layers of short-term memory and of long-term memory.

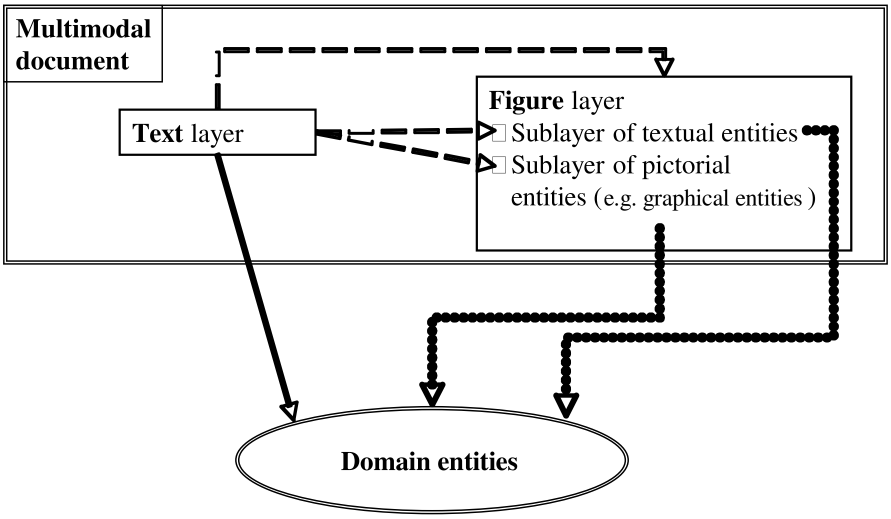

Multimodal document
Text layer
Figure layer
Sublayer of textual entities
Sublayer of pictorial entities (e.g. graphical entities)
Domain entities

Figure 2: Explicit and implicit reference links in multimodal documents (figures with enclosed textual elements).

We will now exemplify the general character of the *reference type* and *representation layer* presented above with text–graphics combinations from the domain of meteorology:[2] referential and co-referential links going out from linguistic entities can consider, firstly, figures as a representational whole, as *‘Figure 3’* in (1), secondly, graphical entities, which are constitutional parts of a figure, as*‘peak’* in (2), and thirdly, entities, which are represented by graphical entities—as ‘warmer period’ in (2) or ‘norm temperature in (1).

(1) *Figure 3* shows the deviation of the average temperature in August from the *norm temperature* w.r.t. the period 1961–1990.

(2) The warmer period starting in the eighties culminates in the peak at 2003.

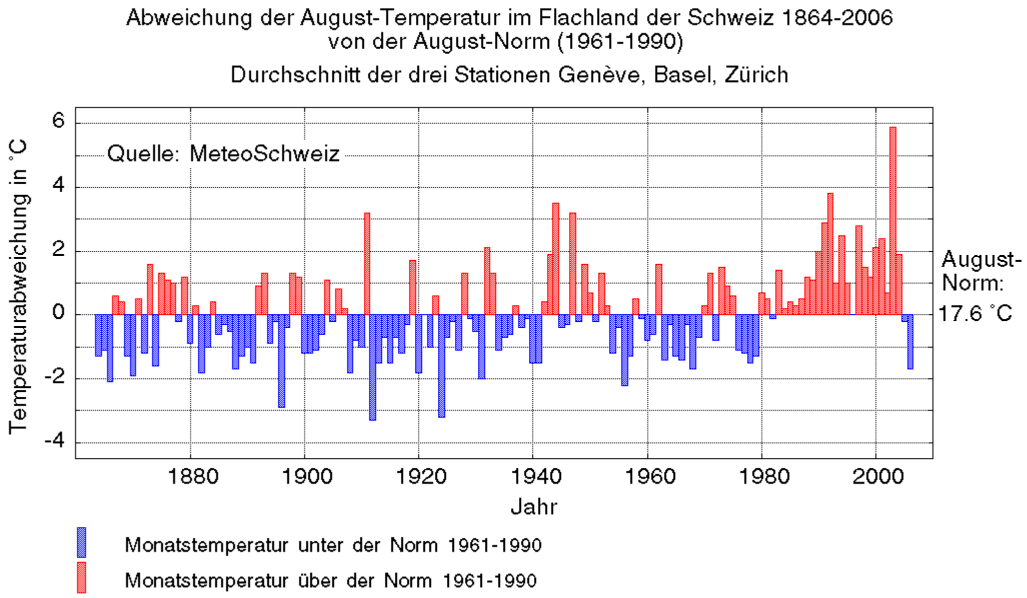


Figure 3: August-temperature chart of Switzerland (©MeteoSwiss / Federal Office of Meteorology and Climatology).

A major difference between maps and information graphics considers the ontological inventory of the domain ‘accessible by pictorial elements’: whereas maps can be seen as—more or less veridical—pictures of geographic space, information graphics are especially used as visualizations of abstract entities, i.e. as externalization of our mental conceptualization of the external world. The domain entities referred to—from sentences (1) and (2) as well as from the information graphics (Figure 3)—have a more or less abstract nature: beyond the geographical frame of Switzerland with some weather station, the temporal dimension of the domain considers ‘calendar entities’ as years and months, periods of temporal entities. Furthermore, the temperature dimension is in the focus of the document, both on the textual and on the graphical layer: the document refers to temperatures, average temperature, deviation of average temperature, etc. (The abstract character of the entities referred to by information graphics will be discussed in more detail in Section 3.)

In comprehending multimodal documents recipients have a range of freedom when to turn to text and when to turn to figures. For example, a figure can get their attention immediately after turning over the

[2] The graphics presented here was published by MeteoSwiss, the Swiss Federal Office of Meteorology and Climatology, on 01.09.2006 as part of a multimodal document with the title “Wie kalt war der August 2006 wirklich?” (How cold was it really during August 2006?). The textual examples used in the present paper are either translations—by the first author—from the original or slight modification to exemplify a phenomenon with respect to the graphics.

page, but the recipients can also neglect focusing on figures until these become relevant to construction of meaning from the text or explicit reference to figures or graphical entities is given verbally. With respect to a particular word or phrase in the text it is difficult to decide which type of reference—to the figure or to the domain—is at issue only by observing the behavior of people comprehending text-figure configurations. For example, 'the peak' in (2) can be interpreted as a *peak in the graph* or as a *peak temperature*, thus resulting in an explicit language-to-graphics reference (to a part of Figure 3) or in an implicit language-to-domain reference via a mediating language-to graphics link. Since the use of definite article in 'the peak' should—according to many approaches to resolving co-reference links—cause the readers to search a peak representation preexisting or easy to find in the discourse model, the two above mentioned alternatives of reference relations can be characterized as 'finding a peak in the accompanying graph' vs. 'constructing (via a bridging inference) a peak in the course of temperature'.[3]

Conversely, producers of multimodal documents have the task to give explicit or implicit hints in the text to lead the recipient's attention as early as necessary to the figure—or even to that parts of the figure—relevant for understanding a sentence or paragraph. Furthermore, producers have to decide in which cases the implicit, mediated language-to-domain reference would not be sufficient and explicit language-to-domain reference has to be realized.

3. IMPROVEMENT: AN ADDITIONAL STEP IN PRODUCING TEXT–GRAPHICS DOCUMENTS

When humans produce multimodal discourse containing language and figures, there exists a spectrum of different grades of coupling the processes of language production, on the one hand, and of the design and realization of figures, on the other hand. Whereas lecturing using chalk and blackboard, e.g. in mathematics or economics—the latter task is used as domain by Tabachneck-Schijf, Leonardo and Simon (1997)—mostly is a one-pass multimodal production process, in which the producer generates speech, writes text and sketches diagrams or graphs in an integrated manner, the production of a newspaper article containing graphics—such as the example discussed in Section 3.1—often is a multi-pass production process, in which the tasks of text production, of graphics production and of text-graphics integration can even be distributed to different people or institutions, in particular, if they possess specific expertise; this case of distributed production corresponds to *type 3* in André's classification (André, 2000, p.309).

Since the dichotomy 'one-pass production process' – 'multi-pass production process' is basic, we will shortly discuss those aspects that are essential for the following. Human production of speech is a prototypical case of a one-pass production process: even repair processes can be seen as part of the only pass (cf. the status of self-perception and self-repair in Levelt's (1999) 'blueprint of the speaker'). When we see co-production of speech and writing or drawing on a blackboard as one task, which can be performed by concurrent processes, then the one-pass perspective is taken with respect to the temporal granularity of speech perception processes. In contrast to this, for humans the production of a route-instruction combining written text and graphics is—on the level of the motor actions of writing and drawing—not realized via concurrent processes. Nevertheless, on the level of designing and realizing a multimodal document, which fulfills specific communicational goals, it is appropriate to see this generation process also as a one-pass process. In particular, since the realizing constraints of humans—on the motoric level—are not relevant for machines, it is appropriate to use one-pass architectures for corresponding computer systems, e.g., the direction services of Mapquest.com. Integrated multimodal generation has been successfully realized using architectures of the 'pipeline' style (Reiter, 1994), or kindred architectures as in WIP (Wahlster et al. 1993) or AutoBrief (Green et al. 2004), in which monitoring and repairing are generation-internal subtasks.

[3] Although cognitive psychology researchers have obtained many fundamental insights in multimodal construction of internal models from text and diagrams (see Glenberg and Langston, 1992; Hegarty and Just, 1993), their models and experiments do seldom focus on the analytical level of the semantics and formal pragmatics of referring expressions, which is in the focus of our research. Thus, we plan empirical investigations—e.g. in the eye-movement paradigm—to support our theoretical models described here.

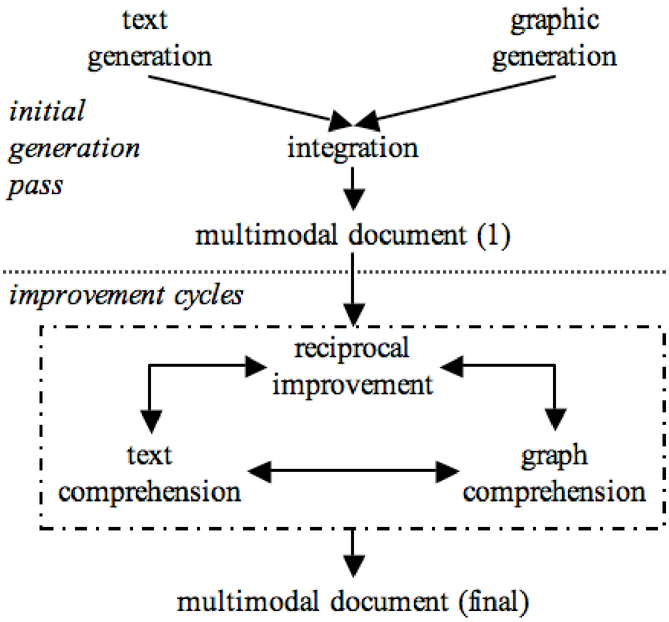


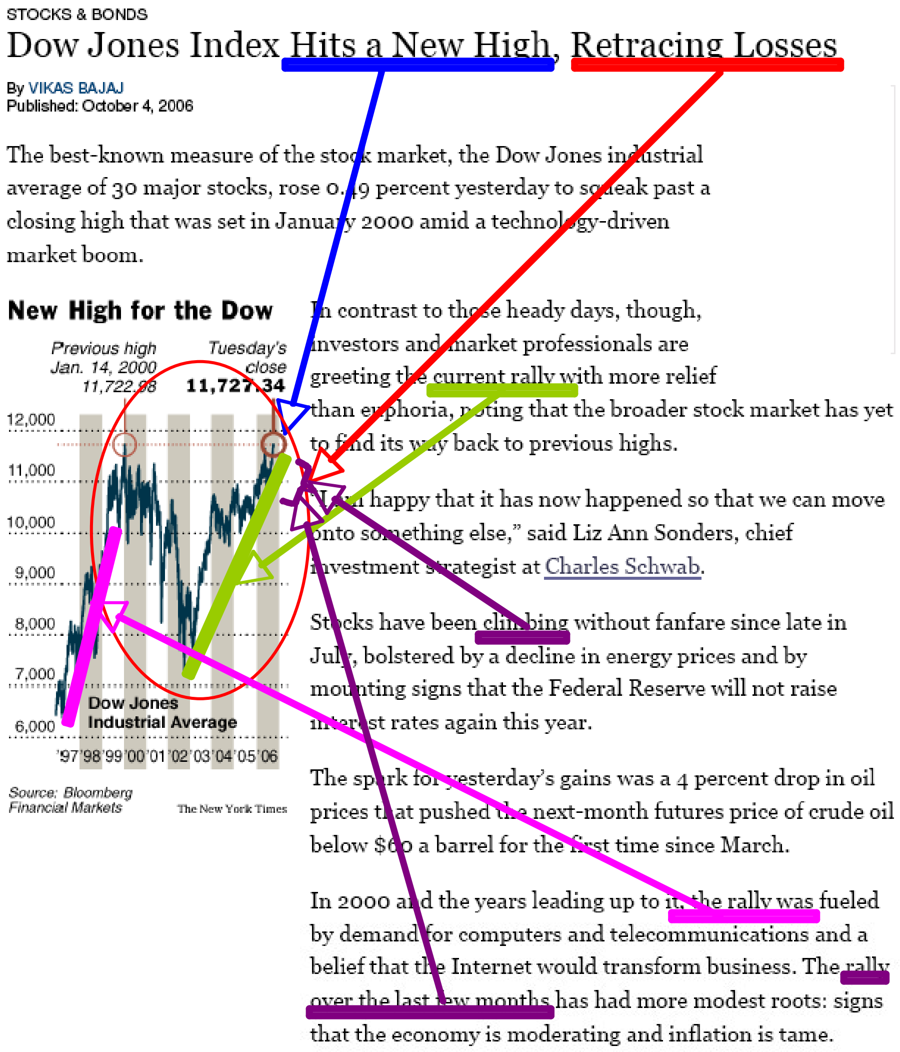
Figure 4: Multi-pass production architecture with improvement module

In contrast, multi-pass production that contains explicit, independent phases of improvement (some authors prefer *revision*) is successful in human text production (cf. Butterfield et al, 1996). Furthermore, Shah, Mayer and Hegarty (1999) show that redesign of graphics increases the quality of viewer's interpretations. Based on these insights from humans' generation of monomodal documents, we propose a multi-pass approach containing an independent 'improvement'—module (see Figure 4) for complex multimodal production tasks. In particular we focus in the present paper on generation, in which modality specific expertise is contributed by loosely coupled, distributed modules or agents.

Improvement of a text-graphic constellation requires comprehension (often called 'interpretation') of both ingredients; in other words, text comprehension and graph comprehension are two basic processes of revising the document (cf. Leinhardt, Zaslavsky and Stein (1990) on the role of interpretation in the construction of graphs). Graph comprehension includes processing of spatial as well as propositional information in different levels by different processes such as preattentive visual processes (Cleveland and Mcgill 1984, 1985, 1987) and interpretive processes (Carpenter and Shah 1998). In multimodal documents, since text comprehension and graph comprehension should lead to an integrated understanding, these processes have to be based on a common system of semantic/conceptual representations, which we will not discuss in detail in the present paper. (Although Green et al. 1998 focus exclusively on the production perspective, their 'content language' is kindred to our representations mentioned above.)

3.1 Improving co-reference by text and information graphics: A case study

In the current section, we discuss multimodal co-reference constellations in an article published in the New York Times on October 4, 2006, to exemplify the requirements on and the tasks of reciprocal improvement in multi-source production of multimodal documents. An excerpt of the text, which was produced by a New York Times author, and the chart provided by Bloomberg Financial Markets, augmented by depictions of referential and co-referential links is depicted in Figure 5. In the following subsection 3.2 we will complement the description of the phenomena with a detailed discussion of the role of conceptual representations in multimodal integration.



STOCKS & BONDS

Dow Jones Index Hits a New High, Retracing Losses

By VIKAS BAJAJ
Published: October 4, 2006

The best-known measure of the stock market, the Dow Jones industrial average of 30 major stocks, rose 0.19 percent yesterday to squeak past a closing high that was set in January 2000 amid a technology-driven market boom.

New High for the Dow
Previous high Jan. 14, 2000 11,722.98
Tuesday's close 11,727.34
12,000
11,000
10,000
9,000
8,000
7,000
6,000
Dow Jones Industrial Average
'97 '98 '99 '00 '01 '02 '03 '04 '05 '06
Source: Bloomberg Financial Markets
The New York Times

In contrast to those heady days, though, investors and market professionals are greeting the current rally with more relief than euphoria, noting that the broader stock market has yet to find its way back to previous highs.

"I am happy that it has now happened so that we can move onto something else," said Liz Ann Sonders, chief investment strategist at Charles Schwab.

Stocks have been climbing without fanfare since late in July, bolstered by a decline in energy prices and by mounting signs that the Federal Reserve will not raise interest rates again this year.

The spark for yesterday's gains was a 4 percent drop in oil prices that pushed the next-month futures price of crude oil below $60 a barrel for the first time since March.

In 2000 and the years leading up to it, the rally was fueled by demand for computers and telecommunications and a belief that the Internet would transform business. The rally over the last few months has had more modest roots: signs that the economy is moderating and inflation is tame.

Figure 5: Dow Jones Index Hits a New High, Retracing Losses, by Vikas Bajaj, published on October 4, 2006 (©The New York Times).

The first part of the article title, i.e. "Dow Jones Index Hits a New High", refers with the object-NP 'New High' either—in the language-to-domain way—explicitly to a domain-entity of the conceptual type '*VALUE OF AN INDEX*'[4] or —in the language-to-graphics way—to a maximum point (graphical entity) of the graph. Nevertheless, the verbal attribute 'new' induces that a 'former high' exists, which has—compared with the 'new high'—only minor salience at this stage of comprehension. On the other hand, in the graph exist two small circles that mark maximum points. To sum up, whereas the former high is only implicitly mentioned in the text—presupposed via the phrase 'New High'—it is explicitly presented in the graphics.

The remaining part of the title, "Retracing Losses" co-refers with a salient V-shaped structure of the graph to a complex two-phase event of '*LOSING*' and subsequent '*RECOVERING FROM A LOSS*'. A detailed interpretation of the graph provides readers with information given neither in the title nor in the further article, namely, about the '*AMOUNT OF THE LOSS*'. In interpreting the Dow Jones Chart, the graph-

[4] We use small italic capitals to denote entities of the conceptual representation layers. According to our prior remarks on the abstractness of some domain entities referred to by text and or information graphics, we characterize in this description abstract domain entities referred to by using the same typographical style as to denote conceptual representations. The role of the conceptual layers in multimodal comprehension and improvement will be discussed in more details in Section 3.2.

comprehension sub-module should detect the V-shaped depiction of the *LOSING & RECOVERING* event in question, which contains the '*DEEP OF 2002*' as a salient protagonist not referred to verbally in the text. The improvement process then could react to this analysis by making changes in one or both modalities, for example, by mentioning the topic '*LOSING & RECOVERING*' verbally either in the title or in a subsequent paragraph, or by editing the V-shape structure in the graph to make the topic more explicit there; furthermore, the improvement process can reject to make any changes.

The second paragraph of the excerpt, (3), includes the phrase 'current rally', which refers to the increase resulted in the second high. The lexeme 'rally' corresponds—in its general meaning—to a process of increasing; in the specialized terminology of stock markets the term refers to a rise or recovery in stock prices. In both cases, 'rally' corresponds to a—co-referring—graph-structure of a specific shape, i.e. the right hand site of the V-shape discussed above. Note that in these comprehension steps different types of knowledge interact: on the one hand, general knowledge about graphs, which is used to detect *gestalts* in graphs, and which is kindred to Pinker's (1990) conception of *graph schemata*, and, on the other hand, knowledge of a sublanguage, namely that of stock markets and charts.

(3) In contrast to those heady days, though, investors and market professionals are greeting the current rally with more relief than euphoria, noting that the broader stock market has yet to find its way back to previous highs.

The third paragraph contains information that cannot be deduced from the graph, i.e. the terms included in this paragraph do not have co-reference relations with the components of the graph, i.e. gestalts in the graph. However, at the beginning of the next paragraph, the focus of the reader is shifted back into the graph: "Stocks have been climbing without fanfare since late in July, ...". Although the term 'climbing' corresponds a specific shape (of the type *INCREASE*), which is similar to the right hand side of a V-shape, the granularity of the figure is poor. As a result, there is no distinguished entity to be identified as a relevant object in the chart. This could prompt the improvement module to modify the chart, either by replacing it with another graph of higher resolution, or by graphical highlighting.

The last paragraph of the excerpt, the first two sentences of which are given in (4), mentions two rallies. The 'rally' in the first sentence refers to sequential increases in year 2000 and previous years. The 'rally' in the following sentence refers to the increase in the rightmost part of the graph, not distinguishable in the graph due to low resolution. Remember that the same '*INCREASE*' was previously referred to in the sentence "Stocks have been climbing without fanfare since late in July, ...". Since the '*RALLY STARTING IN JULY*' becomes now a focused object, it would be appropriate for the improvement process to revise the graphical parts of the document, for example by inserting an additional 6-months-chart.

(4) In 2000 and the years leading up to it, the rally was fueled by demand for computers and telecommunications and a belief that the Internet would transform business. The rally over the last few months has had more modest roots: signs that the economy is moderating and inflation is tame.

3.2 THE CONCEPTUAL LEXICON AS BASIS OF MULTIMODAL INTEGRATION OF TEXT AND GRAPHICS

In particular from the perspective of communication, conceptual representations are the pivot of human cognition. The level of conceptual representations, which encodes meaning independent from any particular language, is the content-specifying level in language comprehension as well as in language production (in psycholinguistics this level in the production process often is called as 'preverbal messages'; cf. Levelt 1999). Furthermore, conceptual structures are an essential part of the interfaces between language and perception as well as action (cf. Jackendoff's interface architecture 1997, 2002), and additionally, conceptual representations are the material for many types of thinking and problem solving. Jackendoff (1997, 2002) uses in his framework the notion 'conceptual structure', which is—by the theoretical and empirical context of his approach—more constrained than the term 'conceptual representation', which has framework-specific variants in interpretation. In the present paper we use these terms as quasi-synonyms, since we cannot specify our framework of conceptual representations for natural language processing in detail here (see Tschander et al. (2002) on CRIL (Conceptual Route Instruction Language), an internal language connecting natural language and action plans, and Guhe et al. (2004) on the use of conceptual representations in language generation).

As we will argue for in this section, conceptual representations play a central role also in multimodal communication. In particular, the interaction between language and graphics is supported by shared conceptual representations. Since conceptual structures are independent from individual languages, the correspondence between lexemes and concepts is usually not of the one-to-one type. In other words, the conceptual counterpart to a lexeme is in most cases a complex structure of conceptual building blocks. The reciprocal assignment of lexemes to conceptual structures and vice versa is fundamental for language comprehension as well as for language production. The relevant knowledge source for these assignments is the lexicon (in Jackendoff's notion, 1997) or the lexical network (using Levelt's terminology, 1999).

In the following, we exemplify the nature of lexeme to conceptual structure relations on a coarse level with the phrase "retracing losses" discussed in Section 3.1. The noun 'loss' (and in a similar manner, the verb 'lose') provides a conceptual representation containing a process concept $DECREASE\_OF\_VALUE(\_{}_{TEMP},\_{}_{VALUE},\ldots)$. We focus here only on two arguments of this process, namely a temporal argument, which can be filled by an interval, and a value argument, which can be filled by an entity of an ordered structure, which functions as the domain of the value. By using such abstract representations, which generalize over different value domains, it is possible to catch the common properties 'loss of money', loss of weight', and others. The temporal argument, which is necessary for all process and event concepts, stands for the 'temporal interval during which the whole process is occurring'. (Note: this does not mean that the producer or the recipient knows how to anchor this interval in physical time.) Putting this together, the process concept *DECREASE_OF_VALUE* stands for a specification of a mapping from the temporal domain in the value domain, or—using the terminology of topology—for a 'path' in the value space.[5] Such abstract topological and geometrical structures are relevant building blocks of conceptual representations in general, not only needed for communication about physical space, but also for types of using what often are called 'figurative language' (cf. Habel, 1990; Habel and Eschenbach, 1997; Eschenbach et al., 1998; Eschenbach et al., 2000).

Let us now look on the lexeme 'retrace', which in some dictionaries is paraphrased as "trace back or trace again". A corresponding concept representation is $INVERSE(\_{}_{PATH})$. The syntactic and semantic analysis of 'retracing losses' specifies that *PATH* corresponding to $DECREASE\_OF\_VALUE(\_{}_{TEMP},\_{}_{VALUE},\ldots)$ is the conceptual argument for $INVERSE(\_{}_{PATH})$. The next step in conceptualizing goes as follows. The inverse following of the value path leads to an $INCREASE\_OF\_VALUE(\_{}_{TEMP},\_{}_{VALUE},\ldots)$ structure. Since a retracing has to occur after the process that is the argument of, the temporal ordering between the two intervals in question, namely the time of the losses and the time of retracing, is inferable: *LOSS* A *RETRACE.* Except from the relation between initial value and final value during the loss-interval, namely *VALUE(BEGIN(LOSS))* > *VALUE(END(LOSS))*, the details of time–value correspondences are not mentioned explicitly in the text. But by a reasonable inference the recipient can infer the following:

$$VALUE(BEGIN(LOSS)) \approx VALUE(END(RETRACE))\ VALUE(END(LOSS)) = VALUE(BEGIN(RETRACE))^{6} \quad (5)$$

In Section 3.1, we proposed a multimodal co-reference constellation constituted by the phrase 'retracing losses' and a V-shaped structure in the chart. At this point of processing the conception of *graph schema* comes into the play (cf. Pinker, 1990; Lohse, 1993): graph schemata provide knowledge to locate and decode information presented in the graph. Whereas Pinker and Lohse focus on the procedural character of graph schemata—i.e. they take a perspective, kindred to Ullman's (1984) *visual routines* approach—we emphasize the abstract spatial properties of graph schemata. Two of the most relevant *gestalt atoms* for line graphs are *increasing* and *decreasing paths*, where 'path' is used as the technical term for (ordered) sequences of line segments. The corresponding entities on the level of conceptual representations are

[5] *Paths* are directed linear entities (cf. Habel, 1990; Eschenbach et al., 2000).

[6] Please remind that we do not specify the conceptual structures in detail in the present paper. Therefore, (5) should be read as 'pseudo-code' and not as formal specification of conceptual meaning.

denoted by $\mathit{INCREASE\_P}(\_{}_{PATH}, \_{}_{SRS})$ and $\mathit{DECREASE\_P}(\_{}_{PATH}, \_{}_{SRS})$ specifying the particular property for a path argument with respect to a 'spatial reference system' (*SRS*). Using an additional concatenation concept $\mathit{CONCAT}(\_{}_{PATH}, \_{}_{PATH}, \_{}_{SRS})$ a V-structure, built by two paths anchored in the same reference system, can be specified by

$$\text{V-STRUCTURE } (\mathit{CONCAT}(\mathit{PATH}_1, \mathit{PATH}_2, \mathit{SRS})) \Leftrightarrow_{\text{def}} \mathit{DECREASE\_P}(\mathit{PATH}_1, \mathit{SRS}) \wedge \mathit{INCREASE\_P}(\mathit{PATH}_2, \mathit{SRS}) \tag{6}$$

Furthermore, the prototypicality of V-structures is determined with respect to additional properties concerning for example the *amount* of increase and decrease, the *order of magnitudes* between these amounts, and their *steepness*. Beyond specific graph schemata, as that of *INCREASE*, *DECREASE* and V-STRUCTURE, there exist more general graph schemata, for example considering the standard *spatial reference system* of line graphs, namely co-ordinate systems with scaled axes, i.e. axes possessing an ordering structure. Exactly the ordering of the y-axes is considered by the graph schema routines, determining what an *INCREASE* and what a *DECREASE* is.

Now, we come back to the process of multimodal comprehension. The phrase 'retracing losses' introduces the internal proxies for two succeeding processes into the discourse model, whose conceptual representations contain *DECREASE_OF_VALUE* and *INCREASE_OF_VALUE* components. Having a complementary line graph in the multimodal document, the graph reader's comprehension component can start a query with respect to a V-structure, corresponding to the characterization (6), which also specifies the correspondence between the conceptual structure built up by language comprehension and the graph schema. The graph schema routines triggered by the abstract spatial description can process the actual graph in goal directed manner; in particular they can be based on using the most salient discourse entities first. In this case, the co-reference of 'new high' and of the induced 'former high' with two small circular entities in the graph should have been built up in the prior step. Thus the 'retracing losses' – graphical V-structure co-reference relation is easy to construct, and additional information can be attached to the discourse units underspecified by the textual information, for example concerning the temporal location of the 'depth', i.e. the point of maximal losses and start of retracing.

4. CONCLUSION AND FUTURE WORK

In the present paper we proposed 'reciprocal improvement' as a means to generate high-quality combinations of text and graphics. These types of revision in later stages of a multi-pass architecture are in particular relevant, if text and graphics are produced by different agents (people, institutions, systems), as in the example we discussed in Section 3.

Since both comprehension modules, which are the base of improvement, namely text comprehension and graph comprehension, employ the same type of conceptual (concept) representation, the lexical and the ontological analysis of terms which refer to 'graphical entities' (e.g., *peak, increase*, etc.) and their verbal encoding in specific domains (e.g., *rally*) are essential. We will perform this analysis via corpus studies as well as by language production experiments with human subjects.

REFERENCES

Ainley, J., Nardi, E. and Pratt, D. (2000). The construction of meanings for trend in active graphing, *International Journal of Computers for Mathematical Learning* 5(2), 85–114.

André, E. (2000). The generation of multimedia presentations. In R. Dale and H. Moisl and H. Somers (eds.), *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. (pp. 305–327). Marcel Dekker Inc.

Beun, R. J. and Cremers, A. H. M. (2001). Multimodal reference to objects: An empirical approach. In H. C. Bunt and R. J. Beun (eds.), *Cooperative Multimodal Communication*. (pp. 64–86). Berlin: Springer-Verlag.

Butterfield, E. C.; Hacker, D. J. and Albertson, L. R. (1996). Environmental, cognitive, and metacognitive influences on text revision: Assessing the evidence. *Educational Psychology Review, 8*. 239-297.

Carpenter, P. A., and Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4(2),* 75-100

Cleveland, W. S., and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 77,* 541–547.

Cleveland,W. S., and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science 229,* 828–833.

Cleveland, W. S. and McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society. Series A (General), 150 (3),* 192-229.

DeProspero, A. and Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*. 573–579.

Eschenbach, C.; Habel, Ch.; Kulik, L. and Leßmöllmann, A. (1998). Shape nouns and shape concepts: A geometry for ‚corner'. In C. Freksa, Ch. Habel and K.F. Wender (eds.), *Spatial Cognition.* (pp. 177–201). Berlin: Springer.

Eschenbach, C.; Tschander, L.; Habel, Ch. and Kulik, L. (2000). Lexical specifications of paths. In C. Freksa, W. Brauer, Ch. Habel and K.F. Wender (eds.), *Spatial Cognition II.* (pp. 127–144). Berlin: Springer.

Glenberg, A. M. and Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language, 31*. 129–151.

Green, N.; Carenini, G.; Kerpedjiev, S.; Mattis, J.; Moore, J. and Roth, S. (2004). AutoBrief: an experimental system for the automatic generation of briefings in integrated text and information graphics. *International Journal of Human Computer Studies, 61*. 32–70.

Green, N.; Carenini, G.; Kerpedjiev, S.; Roth, S. and Moore, J.A. (1998). Media-independent content language for integrated text and graphics generation. *CVIR'98 Content Visualization and Intermedia Representations (*ACL-Coling98 workshop*).*

Guhe, M.; Habel, Ch. and Tschander, L. (2004). Incremental generation of interconnected preverbal messages. In T. Pechmann and C. Habel (eds.), *Multidisciplinary approaches to language production*. (pp. 7–52). Berlin: Mouton de Gruyter.

Habel, Ch. (1990). Propositional and depictorial representations of spatial knowledge: The case of *path* concepts. In R. Studer (ed.): *Natural language and logic*. (pp. 94–117). Lecture Notes in Artificial Intelligence. Berlin: Springer.

Habel, Ch. and Eschenbach, C. (1997). Abstract structures in spatial cognition. In Ch. Freksa, M. Jantzen and R. Valk (Eds.). Foundations of Computer Science – Potential – Theory – Cognition. (pp. 369-378). Springer: Berlin.

Hegarty, M. and Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*. 717–742.

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT-Press.

Jackendoff, R. (2002). *Foundations of language. brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology, 3*. 185–226.

Kosslyn, S. (1994). Elements of graph design. New York: W.H. Freeman.

Kosslyn, S. (2006). Graph Design for the Eye and Mind. OUP.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11.* 65-99.

Leinhardt, G., Zaslavsky, O., and Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*, 1–64.

Levelt, Willem J.M. (1999). Producing spoken language: a blueprint of the speaker. In C.M. Brown and P. Hagoort (eds.), *The neurocognition of language.* (pp. 83–122). Oxford: Oxford University Press.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction, 8*, 353–388.

Peebles, D. J., and Cheng, P. C.-H. (2001). Graph-based reasoning: from task analysis to cognitive explanation. In J. D. Moore and K. Stenning (Eds.), *Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society* (pp. 762-767). Mahwah, NJ: Lawrence Erbaum.

Pinker, S. (1990). A theory of graph comprehension. In R.O. Freedle (ed.), *Artificial intelligence and the future of testing.* (pp. 73–126). Hillsdale, NJ: Erlbaum.

Publication Manual of the American Psychological Association (4th ed.). (1994). Washington, DC: American Psychological Association.

Reiter E. (1994) Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? *IWNLG-1994*, 163–170, Kennebunkport, ME.

Shah, P., Mayer, R. E., and Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, 91(4), 690-702.

Tabachneck-Schijf, H. J. M.; Leonardo, A. M. and Simon, H. A. (1997). CaMeRa: A Computational Model of Multiple Representations. *Cognitive Science, 21.* 305–350.

Tappe, H. and Habel, Ch. (1998). Verbalization of dynamic sketch maps: Layers of representation and their interaction. [Full version of one page abstract / poster at Cognitive Science Conference; Madison WI, August, 1.-4., 1998.] ftp://ftp.informatik.uni-hamburg.de/pub/unihh/informatik/WSV/TappeHabel_CogSci_1998.pdf

Tschander, L.; Schmidtke, H.; Habel, Ch.; Eschenbach, C. and Kulik, L. (2003). A geometric agent following route instructions. In Ch. Freksa, W. Brauer, Ch. Habel and K. Wender (eds.), *Spatial Cognition III.* (pp. 89–111). Berlin: Springer.

Ullman, S. (1984). Visual routines. *Cognition, 18.* 97–159.

Wahlster, W.; André, E.; Finkler, W.; Profitlich, H. -J. and Rist, Th. (1993). Plan-based integration of natural language and graphics generation. *Artificial Intelligence, 63.* 387–427.

Winn, W. (1991). Learning from maps and diagrams. *Educational Psychology Review*, 3, 211–247.

Multimodal Backchannel Generation for Conversational Agents

Dirk Heylen*
Human Media Interaction - University of Twente
PO BOX 217, 7500 AE Enschede, The Netherlands
heylen@ewi.utwente.nl

Abstract

Listeners in face-to-face interactions are not only attending to the communicative signals being emitted by the speakers, but are sending out signals themselves in the various modalities that are available to them: facial expressions, gestures, head movements and speech. These communicative signals, operating in the so-called back-channel, mostly function as feedback on the actions of the speaker; providing information on the reception of the signals; propelling the interaction forward, marking understanding, or providing insight into the attitudes and emotions that the speech gives rise to.

In order to be able to generate appropriate behaviours for a conversational agent in response to the speech of a human interlocutor we need a better understanding of the kinds of behaviours displayed, their timing, determinants, and their effects. A major challenge in generating responsive behaviours, however, is real-time interpretation, as responses in the back-channel are generally very fast. The solution to this problem has been to rely on surface level cues. We discuss on-going work on a sensitive artificial listening agent that tries to accomplish this attentive listening behaviour.

Keywords: listener responses, backchannels, head movements

1 Introduction

Schegloff (1982) qualifies face-to-face interaction as an interactional achievement. Communication is constituted by the collaborative action of multiple actors. Gumpertz (1982) states this as follows.

> Communication is a social activity requiring the coordinated efforts of two or more individuals. Mere talk to produce sentences, no matter how well formed or elegant the outcome, does not by itself consitute communication. Only when a move has elicited a response can we say communication is taking place.

Responses to talk by recipients can take many forms. They can take the form of a subsequent move in the next turn, but most of the behaviors of the non-talking participant in the conversation displayed during the turn of the speaker can count as some kind of "response", providing the speaker with feedback on perception, attention, understanding and the way in which the message is received in general: the change in the beliefs, attitudes and affective state of the recipient. These cues and signals enable the synchronization of the communicative actions (for instance turn-taking), grounding and the building of rapport.

Most of the work on the generation of communicative behaviors of embodied conversational agents has been concerned with generating the appropriate non-verbal behaviors that accompany the speech of the embodied agent: the brows, the gestures, or the lip movements. The generation

*The author would like to thank Ruben Kooijman and Boris Reuderink for their work on this project. This research is part of a wider project on listening behaviours for conversational agents in the context of the Humaine Network of Excellence in collaboration mainly with Catherine Pelachaud and Elisabetta Bevacqua (Paris).

of the verbal and non-verbal behaviors to display during the production of speech by another actor, that is the behavior of a listening agent, has received less attention. A major reason for this neglect is the inability of the interpretation modules to construct representations of meaning incrementally and in real-time, that is contingent with the production of the speech of the interlocutor. As many conversational analists and other researchers of face-to-face interaction have shown, the behaviors displayed by auditors is an essential determinant of the way in which conversations proceed. By showing displays of attention, interest, understanding, compassion, or the reverse, the auditor/listener, determines to an important extent the flow of conversation, providing feedback on several levels.

Besides the fact that most work on embodied conversational agents has focused on speaking behaviors, it also appears that not all expressive behaviors have received the same amount of attention. Language, facial expressions, gestures and gaze are the main kinds of expressive behaviors that have been studied so far. Posture and head movements form another group of nonverbal behaviours, that are very informative about the intentions, attitudes, emotions and the mental state of interlocutors, in particular, “auditors”, but these have been less widely studied.

In our current work on the Sensitive Artificial Listener, the generation of the behaviours that an agent should display while listening are very important. In our first studies we are looking at head movements and gaze in particular. In this paper we describe the general contours of the project, the way we approach the subject and illustrate this with describing the set-up and results of a pilot experiment.

2 Sensitive Artificial Listener

In the Sensitive Artificial Listening Agent project, we are attempting to build semi-autonomous embodied chat-bots as part of the Sensitive Artificial Listener software. This software, developed in collaboration with Queens University, Belfast (see http://www.emotion-net.research/), is used to elicit emotions and accompanying behaviours that occur in conversations. In the original system, a person is sitting in front of a camera and hears the voice of one of the “characters”. The utterances by the characters are selected by an operator who can choose from a collection of pre-recorded phrases. They are indexed by the character they belong to, a pair of emotion dimension labels (positive/negative and active/passive) and by content category. They consist of general moves such as greetings, questions that prompt the persons interacting with SAL to continue speaking, and all kinds of reactions to what the persons are saying. The operator chooses the particular utterances in accordance with the stage of the dialogue and the emotional state of the person. Each character has a different personality expressed through what they are saying and the way they say it and will try to bring the person in a particular emotional state by their comments; cheerful or gloomy, for instance. In the Agent version that we are developing, the voices are replaced by talking heads and the behaviours are partly decided upon automatically.

We are working on the following items.

1. Designing faces and animations that fit the different personalities of the characters
2. Deciding on animations of the nonverbal behaviours that the characters should display when uttering the canned phrases
3. Studying and implementing the behaviours that should be displayed while listening
4. Building a system with some perception and understanding capabilities
5. Building a system that can decide semi-autonomously on which behaviours to display

Our work on building a completely autonomous responsive listening agent is proceeding by making small steps at the time. A detailed data-analysis is needed as the variations and functions of backchannels are highly varied. The real-time detection of features to which the agent can respond is in need of a better understanding of the function of those features in different contexts. Besides the need to know whether a backchannel has to be generated, it is important as well which kind of backchannel is called for to what effect. Each of these issues can be further investigated in various ways. Besides data analysis, we are also relying on perception experiments in which people have to rate the behaviours of embodied agents in various ways. One of the first experiments that we have carried out in this respect will be discussed in Section 5.

Before we can experiment with fully autonomous agents, some other versions are being tried out. We have started with the off-line simulation of listening behaviours in which we generate behaviours for an agent and show these in combination with pre-recorded fragments of humans to make it appear as if the human interlocutor was talking to the agent. The behaviours are either generated by hand or by rule. In the latter case we are experimenting with probabilistic models and rule based systems. The road to full autonomy takes the following steps:

1. Off-line simulation of listening behaviours
2. Wizard of Oz experiments
3. Experiments with semi-autonomous agents

In this paper we will give an overview of the objectives and the way we approach the project. We discuss the collection and analysis of data, the implementation of head trackers and modules that drive the behaviour of the agent on the basis of rules and regularities found during data collection and some initial experiments on the evaluation of some model implementations. We start by providing some background on the phenomena under investigation.

3 Background

Several research traditions have studied the behaviours that listeners display in conversations. Back-channels, or similar phenomena with a different name such as response tokens, have been studied in the conversational analysis literature, for instance, with the purpose of understanding what role the various contributions of all of the participants play in shaping the conversation. Most relevant in this respect are papers such as Schegloff (1982), Schegloff (1996), Heritage (1984) but there are many others. The literature on turn-taking, both from the CA and other perspectives, also provides useful notes on the behaviours of participants that assume the primary speaker role and the auditors. In the series of papers by Duncan and co-authors[1], for instance, auditor back-channel signal are one of three classes of signals, besides speaker within-turn and speaker continuation signals, that serve to mark units of interaction during speaking turns.

An important issue that comes up with the study of back-channels is the definition of such terms as *speaker*, *hearer* and synonyms. A general assumption behind the concept of back-channel is that all the participants in a face-to-face conversation are both producers and recipients of communicative signals, but that there are different levels on which this occurs. Communicative signals on the primary track, to use the term by Clark (1996), are by the participants that have the floor and the secondary track, 'in the back', is constituted by the feedback on the behaviours in the primary track. As Yngve (1970) points out there may be cases of iteration where speakers provide feedback on the back-channels of listeners. To make the definitions of back-channel more precise, one would therefore need a framework that describes the various roles participants take in interaction. We build on the work of Goffman (1981), Levinson (1988), Schegloff (1996), Clark (1992) as a starting point for our theoretical model. However, in this paper, we will be using the terms speaker and hearer without further concerns about the tricky issues that surround them.

Several studies of nonverbal behaviours have paid attention to the behaviours displayed by listeners. One kind of phenomenon that has received some attention is the way in which behaviours of participants are synchronized and in particular how body movements of listeners are coordinated with the verbal utterances of the speaker. Hadar et al. (1985) showed that about a quarter of the head movements by listeners are in sync with the speaker's speech. Interactional synchrony in this sense has been studied, amongst others by Kendon (1970), Scheflen (1964), Condon and Ogston (1967). Mirroring is a particular type that has often been commented upon. Scheflen suggests

[1] See Duncan (1972), Duncan (1973), Duncan (1974), Duncan (1976), Duncan and Niederehe (1974),.

that this often reflects a shared viewpoint. Also Kendon (1970) hypothesized that the level to which behaviours are synchronized may signal the degree of understanding, agreement or support. These kinds of phenomena show that the behaviours of listeners arise not only from 'structural concerns' (e.g. turn-taking signals) but also from 'ritual concerns'. We take these terms from Goffman (1981) who points out that it is sheer impossible to assign to behaviours a function of only one of these types of concerns (see also Bernieri (1999)).

Besides these synchrony behaviours, listeners display various other nonverbal behaviours as feedback. Chovil (1991), looking in particular at facial expressions, classifies these behaviours in a small set of semantic categories of listener comment displays. These are, besides displays for agreement:

- Back-channel: displays that were produced by listeners while the speaker was talking or at the end of the speaker's turn. They take the form of brow raises, mouth corners turned down, eyes closed, lips pressed. In Chovil's corpus the displays could be accompanied by typical back-channel vocalizations such as "uhuh", "mhmm", "yeah", etc.
- Personal reaction displays: a reaction in response to what the speaker had said rather than just acknowledging the content.
- Motor mimicry displays: displays that might occur in the actual situation that the speaker is talking about (e.g. wincing after hitting ones' thumb with a hammer, eyes widened and an open mouth in response to a frightening situation). These are interpreted as messages that indicated a sincere appreciation of the situation being described.

Hadar and colleagues have looked in particular at head movements of listeners and how differences in form correspond to functional differences. Several authors writing on head movements have remarked that the precise form of the movements may be informative about the different functions they serve. Kendon (2003), writing on head shakes for instance, states:

> Head shakes vary in terms of the amplitude of the head rotations employed, in the number of rotations and in the speed with which they are performed. There is no doubt that these variations in performance intersect with and modify the meaning of the gesture. [...] In this paper, however, necessarily preliminary as it is in many ways, we have made no attempt to subdivide the head shake according to how it may be varied in its performance [...].

In Hadar et al. (1985), such an attempt has been made for a limited number of head movements. They show that kinematic properties such as amplitude, frequency and cyclicity distinguish between signals of 'yes and no' (symmetrical, cyclic movements), anticipated claims for speaking (linear, wide movements), synchrony movements occurring in phase with stressed syllables in the other's speech (narrow, linear) and movements during pauses (wide, linear). As we shall illustrate below, we are performing work along similar lines to reach a better understanding of how the variations in form give rise to variations in meaning.

In the discussion so far, we have mentioned several functions that are served by the behaviours of listeners. They provide feedback to the speaker, acknowledging reception of the signal, possibly its understanding or some kind of comment expressing a particular attitude towards what is being expressed. From its nature as a kind of joint communicative action, conversations require that participants come to react to each other's actions to ground the actions and provide closure. Feedback is an important part of establishing grounding in the interactional achievement of having a conversation. The variety of functions that feedback serves is partly explained by the various levels on which grounding needs to take place: i.e. levels at which the participants need to have a mutual understanding of each other's intentions. Clark (1996) suggests that grounding needs to occur on at least four levels with each step a kind of joint action.

1. Joint[A executes behavior t for B to perceive; B attends perceptually to behavior t from A]

2. Joint[A presents signal s to B, B identifies signal s from A]

3. Joint[A signals to B that p, B recognizes that A means that p]

4. Joint[A proposes a joint project to B, B takes up the joint project]

As speakers make their utterances, they are usually also monitoring the interlocutors behaviours to find signs of their participatory involvedness on all of these levels.

1. A monitors B for signs of perception activity / B's behaviour provides cues of perception activity

2. A monitors B for signs that B has identified the signal / B indicates that he has identified the signal...

The utterance of speakers and the accompanying behaviours will often be designed to invoke behaviours of interlocutors to ensure this. A typical case of this behaviour is analysed by Goodwin (1981), consisting of hesitations and repetitions of speakers at the beginning of their utterance to evoke gaze behaviours in interlocutors.

In a similar vein, Allwood et al. (1993) distinguishes four basic communicative functions on which the speaker may require feedback:

1. Contact: is the interlocutor willing and able to continue the interaction

2. Perception: is the interlocutor willing and able to perceive the message

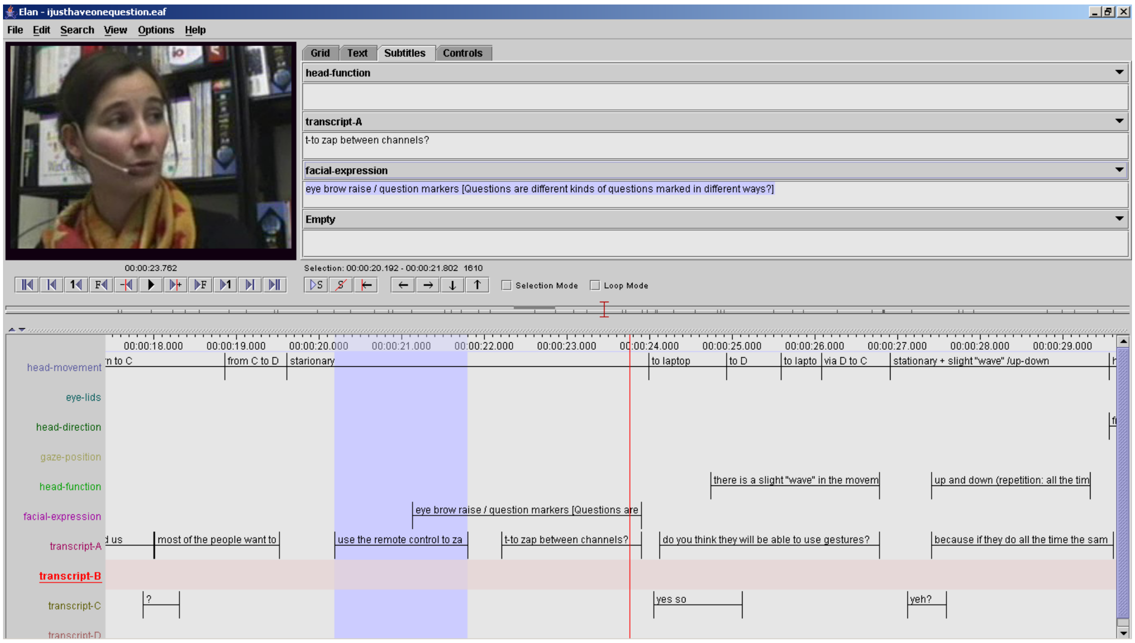
3. Understanding: is the interlocutor willing and able to understand the message

4. Attitude: is the interlocutor willing and able to react and respond to the message (specifically accepting or rejecting it).

The various feedback behaviours are thus not only varied in their form but also in their function. The timing of them is of the essence, as several forms occur in parallel with the utterance of the speaker (synchronous interaction). This poses a big challenge for constructing embodied agents that need to react instantly on the speech produced by speakers. Most of the work on reactive agents has based the reactions on superficial cues that are easy to detect. The listening agent developed at ICT (Maatman et al. (2005) and Gratch et al. (1996)) produces feedback on the basis of head movements of the speaker and a few acoustic features (Ward and Tsukahara (2000)). Similar kinds of input will be used in the SAL system.

4 Data

Although there is a fairly rich literature on listener behaviours that be can used to define and implement behaviours of a listening agent, we have also found it useful to look at some of the data that we have available and to collect some new data to learn more about the form, the distribution and the functions of feedback behaviours. Both the corpus collected in the AMI project (http://www.amiproject.org/) and the collection of interactions with the Wizard of Oz version of SAL (see http://www.emotion-research.net/) are being used in this respect.



head-function:

transcript-A: t-to zap between channels?

facial-expression: eye brow raise / question markers [Questions are different kinds of questions marked in different ways?]

Empty:

Selection: 00:00:20.192 - 00:00:21.802 1610

head-movement: n to C | from C to D | starionary | to laptop | to D | to lapto | via D to C | stationary + slight "wave" /up-down

eye-lids:

head-direction:

gaze-position:

head-function: there is a slight "wave" in the movem | up and down (repetition: all the tim

facial-expression: eye brow raise / question markers [Questions are

transcript-A: d us | most of the people want to | use the remote control to za | t-to zap between channels? | do you think they will be able to use gestures? | because if they do all the time the sam

transcript-B:

transcript-C: ? | yes so | yeh?

transcript-D

The screenshot shows some of the annotations that have been made on the AMI data. In this case, the annotations covered head movements, gaze position, facial expressions, a characterisation of the function of the head movements and the transcripts. The characterisation of the functions was based on the determinants listed in Heylen (2006), which relied in their turn on some of the literature cited above, in particular Chovil (1991) and McClave (2000). The following lists provides the major functions that were used for head movements and back-channels.

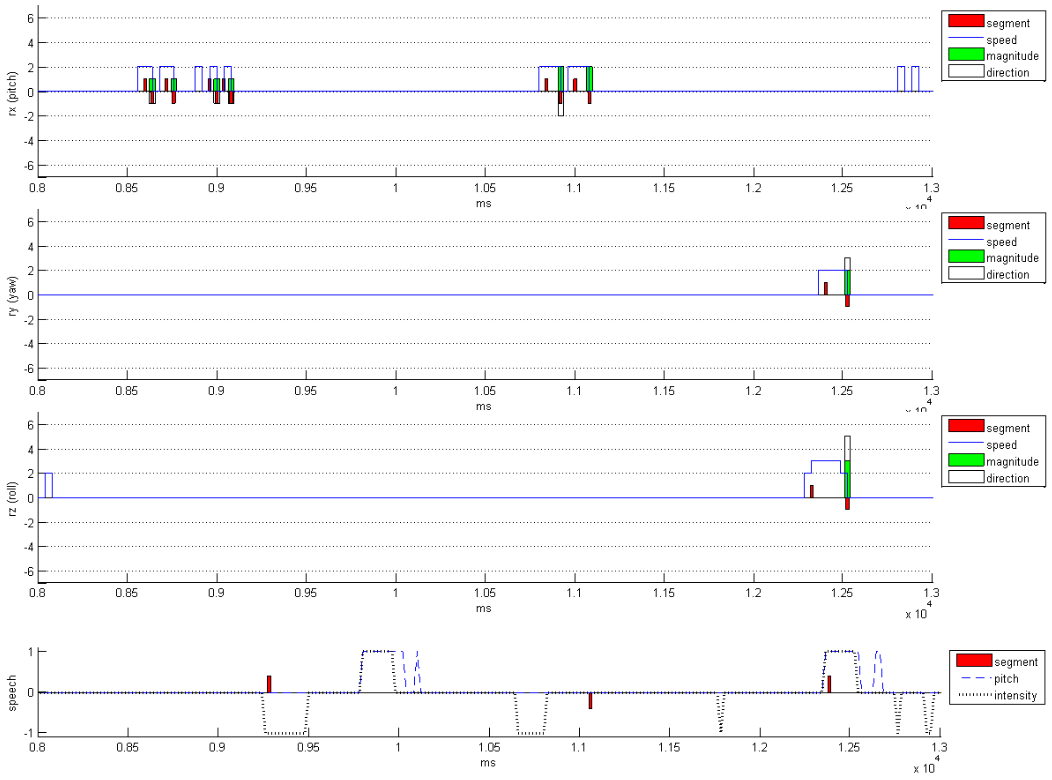
1. Cognitive determinants: thinking, remembering, hesitation, correction...

2. Interaction management: turn-regulation...

3. Discourse and information functions: deixis, rhetorical functions (including the narrative functions mentioned in McClave (2000)), question marker, emphasis...

4. Affect and Attitude: epistemic markers (belief, scepticism), surprise, etcetera.

Besides such hand-made annotations, we are using automatic procedures to extract features from the speech and movements (head movements) to be able to list in more detail the distribution of verbal and nonverbal backchannels. The following picture provides a graph representing the head movements of a particular fragment in the SAL data. It shows the various movements of the head on the x, y and z axes in a small fragment of time with an indication of the duration, the velocity and the amplitude.



For this particular fragment the speech runs as “[lip smack]...euhm...well”. Just before the lip smack the head turns slightly upwards as can be read from the first markers on the pitch line. One set of annotations produced for the first half of this fragment is as follows. The form table gives a brief specification of the various behaviours.

Form		
Head Eyes Speech Face	tilted left right corners looking right/up lip smack + ‘euhm’ anxious	movement up small shakes

The “Intention” values give an indication of the degree to which the head movement was judged to be deliberate or automatic. In many cases both will apply to some extent.

Function	
Intention	deliberate / automatic
Cognition	thinking
Attitude	uncertainty
Social convention	
Emotion	neutral / anxious
Information	
Interaction	stall - responsive - turn-initialisation
Discourse	

With an analysis of some hundred fragments like these we hope to get a better picture of the association between head movements properties and their functions, where the head movements can be quite tiny and the functions more refined than in most analyses currently available in the literature.

5 Impression management

The personality of the four characters used in SAL comes out mainly in the kinds of things they say. The character Poppy, for instance, is cheerful and optimistic and will try to cheer up the interlocutors when they are in a negative state and be happy for them when they are in a positive state. Obadiah, on the other hand, is gloomy and passive and will say things with the opposite effect. The voices, created by (amateur) actors are also quite expressive. The choice of talking head should match this, as should their nonverbal behaviors. A good deal of this work might be left to an animator who is skilled in designing and animating characters. An important question with respect to evaluation in this case is what impression the characters generate. So far, we haven't put animators to work on creating particular animations, but we have carried out some experiments varying the gaze behaviour and the head movements of a character and having participants in the experiment judge these behaviours. The basis of our study were the experiments carried out earlier by Fukayama et al. (2002) for gaze and Mignault and Chauduri (2003) for head movements, complemented by many other works on gaze[2] and head movements.

Similar to the study in Fukayama et al. (2002), a probabilistic model of the behaviours was implemented that determined the gaze of the RUTH talking head (Reuderink (2006)). We limited the variation in movements by fixing the head tilt. Combining some of the outcomes of the two studies we tried to model the behaviours for a happy, friendly, unobtrusive, extrovert agent (A) and for an unhappy, unfriendly and rather dominant agent (B). The combination of two different behaviours together with the fact that different impression variables were attempted to be modeled, raised some interesting issues. The head tilt for A was set to +10° (raised). According to the study by Mignault and Chauduri a head tilted upwards can be perceived as more dominant which is not exactly what we wanted, but it also has an effect on the impression of happiness, which is what we aimed for.

For A, the amount of gaze was set at 75% and a short mean gaze duration which we hoped would create the impression of engagement, friendliness and liking. The mean gaze duration for A was set at 500ms as in the Fukayama et al. (2002) experiment short gaze durations were associated with friendly characters. Gaze aversion for A was downwards, which is associated with submissive rather than dominant personalities.

For B the head tilt was 0°, which may lead, according to Mignault and Chauduri to low scores on happiness. With respect to gaze, we kept the amount of gaze at 75% but changed the mean gaze duration to 2000ms, which results in long periods of gaze, which we hoped would create a rather dominant, unfriendly impression. Gaze aversion for B was to the right.

The settings for both characters are summarised in the following table.

A	
Personality:	happy, friendly, unobtrusive, extrovert
Head tilt	10°
Amount of gaze	75%
Mean gaze duration	500ms
Gaze aversion	down
B	
Personality	unhappy, unfriendly, dominant
Head tilt	010°
Amount of gaze	75%
Mean gaze duration	2000ms
Gaze aversion	to the right

For both A and B we made two animations, one with smaller (A1, B1) and one with larger movements (A2, B2). Each animation we showed in the experiment lasted 40 seconds. We showed the four movies to 21 participants (all students at the University of Twente), divided into three groups for each of which the movies were presented in a different order (A1 B1 A2 B2; B1 A1 B2 A2; A2 B2 A1 B1). The difference in ordering did not show an effect on the result. To rate the

[2] Argyle and Cook (1976),Cassell and Thórisson (1999), Kendon (1967), and our own work Heylen et al. (2005).

impressions we had the participants fill out a questionnaire for each movie consisting of a rating on a 7-point scale for 39 dutch adjective pairs with the following translations.

> extrovert - introvert, stiff - smooth, static - dynamic, agitated - calm, closed - open, tense - relaxed, sensitive - insensitive, polite - rude, suspicious - trusting, interersted - uninterested, credbile - incredible, sympathetic - unsympathetic, self-confident - uncertain, cold - warm, weak - strong, selfish - compassionate, formal - informal, winner - loser, thougthful - reckless, unattractive - attractive, organized - disorganized, unfriendly - friendly, reliable - unreliable, refined - rude, involved - distant, flexible - linear, amusing - boring, attentive - absent, lazy - industrious, inactivy - lively, optimistic - pessimistic, happy - depressed, loving - unloving, empathetic - unempathetic, dominant - submissive, aggressive - timid, stubborn - willing, enterprising - passive, realistic - artificial.

Factor analysis reduced the number of dimensions to the following 8 factors.

1. absence, unfriendliness, rudeness
2. submissive, weak, sensitive
3. warm, energetic
4. dull, drained
5. unreliable
6. rigid, static, linear
7. informal
8. attractive

When A and B are compared on these factors, we found that A scores higher on Factors 2 (submissiveness) and 5 (unreliability). B scores significantly higher on Factors 1 (absence, unfriendliness...) and 4 (dullness). There are also some differences between the small and large movements. Large movements create a more unfriendly impression (Factor 1). Small movements score significantly higher on Factor 2 and 8, that is, the smaller movement animations are considered more submissive, but are also more attractive.

All in all we were thus able to generate behaviours that resulted in several impression values that we had designed the agents for. However, this way of designing behaviours by combining functions associated with behaviours as mentioned in the literature poses many interesting problems. As we mentioned before, if one tries to achieve an effect on various impression variables, a particular behaviour may be very well suited for yielding good scores on variable x but mediocre scores on variable y. Also the combination of two behaviours may yield a combined effect that is different from what might be expected from the descriptions of the behaviours independently considered. Adding yet more behaviours - such as speech, for instance - may change the results again.

Furthermore, the precise set of impression categories that one is aiming for may not correspond exactly to the categories used in the studies in the literature. Expressions are ambiguous and fit more than one category.

Context plays an important role as well. The literature reports on functions and impressions, derived from data in a particular context which can be very different from the context of use that we are considering. For the actual design of the style of behaviours for the different personalities we will rely on another methodology. Using actors to collect a corpus of behaviours would be a good option in this case. It will be interesting then to see what gaze behaviours and head movements they display and see how this compares to the literature and the results of this more analytic approach exemplified by the current study. Despite the many obstacles this kind of approach poses, it does produce some useful results as well.

6 Conclusion

The SAL context provides us with an interesting set-up to experiment with designing and implementing conversational agents. The fact that the system is in part a wizard of oz set-up makes it possible to have the operator make decisions that need high-level interpretation. Because the agents are primarily designed to "listen" it is important to look in more detail at these less well-studied behaviours.

In this paper, we have presented the way we are proceeding to tackle this project. We have illustrated the process of data collection, analysis and one of the ways in which we are evaluating the generation of multimodal listening behaviours. We believe that a project such as this one can only succeed if many different sources and methodologies are brought into play as the above will have shown.

References

Allwood, J., Nivre, J., and Ahlsén, E. (1993). On the semantics and pragmatics of linguistic feedback. *Semantics*, 9(1).

Argyle, M. and Cook, M. (1976). *Gaze and Mutual gaze.* Cambridge University Press.

Bernieri, J. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal behavior*, 23(4):253–269.

Cassell, J. and Thórisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538.

Chovil, N. (1991). Social determinants of facial displays. *Journal of Nonverbal Behavior*, 15(3):141–154.

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

Clark, H. H. (1992). *Arenas of Language Use.* The University of Chicago Press, Chicago, London.

Condon, W. and Ogston, W. (1967). A segmentation of behavior. *Journal of Psychiatry*, 5:221–235.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–92.

Duncan, S. (1973). Towards a grammar for dyadic conversations. *Semiotica*, pages 29–46.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 2:161–180.

Duncan, S. (1976). Language, paralanguage, and body motion in the structure of conversations. In McCormack, W. and Wurm, S., editors, *Language and Man. Anthropological Issues*, pages 239–268. Mouton, The Hague.

Duncan, S. and Niederehe, G. (1974). On signalling that its your turn to speak. *Journal of Experimental Social Psychology*, 10:234–47.

Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., and Hagita, N. (2002). Messages embedded in gaze of interface agents - impression management with agent's gaze. In *Proceedings of CHI 2002*, pages 41–48. ACM.

Goffman, E. (1981). *Forms of Talk.* Oxford University Press, Oxford.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers.* Academic Press, New York.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., and Morency, L.-P. (1996). Virtual rapport. In Gratch, J., Young, M., Aylett, R., Ballin, D., and Olivier, P., editors, *Intelligent Virtual Agents*, pages 14–27.

Gumpertz, J. (1982). *Discourse Strategies*. Cambridege University Press, Cambridge.

Hadar, U., Steiner, T., and Rose, C. F. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In Atkingson, J. M. and Heritage, J., editors, *Structures of Social Action*. Cambridge University Press, Cambridge.

Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(3):241–267.

Heylen, D., van Es, I., van Dijk, B., and Nijholt, A. (2005). Experimenting with the gaze of a conversational agent. In van Kuppevelt, J., Dybkjaer, L., and Bernsen, N. O., editors, *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. Kluwer Academic Publishers.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63.

Kendon, A. (1970). Movement coordination in social interaction: some examples described. *Acta Psychologica*, 32:100–125.

Kendon, A. (2003). Some uses of head shake. *Gesture*, 2:147–182.

Levinson, S. C. (1988). Putting linguistics on a proper footing: explorations in goffman's concept of participation. In Drew, P. and Wootton, A., editors, *Erving Goffman. Exploring the Interaction Order*, pages 161–227. Polity Press, Cambridge.

Maatman, R., Gratch, J., and Marsella, S. (2005). Natural behavior of a listening agent. In *5th International Conference on Interactive Virtual Agents*. Kos, Greece.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

Mignault, A. and Chauduri, A. (2003). The many faces of a neutral face: Heat tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*, 27(2):111–132.

Reuderink, B. (2006). The influence of gaze and head tilt on the impression of listening agents.

Scheflen, A. (1964). The significance of posture in communication systems. *Psychiatry*, 27:316–331.

Schegloff, E. A. (1982). Discourse as interactional achievement: Some uses of "uh huh" and other things that come between sentences. In Tannen, D., editor, *Analyzing discourse, text, and talk*, pages 71–93. Georgetown University Press, Washington, DC.

Schegloff, E. A. (1996). Issues of relevance for discourse analysis: Contingency in action, interaction and co-participant context. In Hovy, E. H. and Scott, D. R., editors, *Computational and Conversational Discourse. Burning issues - An interdisciplinary account*, pages 3–35. Springer.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanes. *Journal of Pragmatics*, 23:1177–1207.

Yngve, V. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pages 567–77, Chicago: Chicago Linguistic Society.

Towards Automatic Generation of Multimodal Answers to Medical Questions: A Cognitive Engineering Approach

Charlotte van Hooijdonk, Emiel Krahmer, Alfons Maes
Tilburg University
Tilburg, The Netherlands
{c.m.j.vanhooijdonk | e.j.krahmer | maes}@uvt.nl

Mariët Theune, Wauter Bosma
University of Twente
Enschede, The Netherlands
{m.theune | w.e.bosma}@utwente.nl

Abstract

This paper describes a production experiment carried out to determine which modalities people choose to answer different types of questions. In this experiment participants had to create (multimodal) presentations of answers to general medical questions. The collected answer presentations were coded on types of manipulations (typographic, spatial, graphical), presence of visual media (i.e., photos, graphics, and animations), functions and position of these visual media. The results of a first analysis indicated that participants presented the information in a multimodal way. Moreover, significant differences were found in the information presentation of different answer and question types.

Keywords: Multimodal information presentation, cognitive engineering

1 INTRODUCTION

Much research on question answering (QA) has focussed on answering factoid questions, i.e., questions that have one word or phrase as their answer, such as "Amsterdam" in response to the question "What is the capital of the Netherlands?" Obviously, factoid QA does not really require Natural Language Generation, and the output modality will typically be text. However, there is currently a growing interest in moving beyond factoid questions and purely textual answers, and then output generation becomes an important issue. Questions that arise are: how to determine for a given question, what the best combination of modalities for the answer is? And related to this: what is the proper length of a non-factoid answer? In this paper, we describe ongoing work in the context of a medical QA system within the IMIX / IMOGEN[1] project that addresses exactly these issues.

In the medical domain several question types occur, such as definition questions or procedural questions. These different types of questions require different types of answers. For example the answer to the definition question "What does RSI stand for?" would probably be a textual answer, like "RSI stands for Repetitive Strain Injury". However, the presentation of an answer through text only may not be the best choice for every type of information. In some cases other modalities (e.g., pictures, film clips, etc.) or modality combinations (e.g., text + picture) may be more suitable. For example the answer to the procedural question "How should I organize my workspace in order to prevent RSI?" would probably be more informative if it contained a picture. Moreover, the length of the answer could also play an important role in the answer presentation. For example, the answer to the question "What does RSI stand for?" could be an extended one: "RSI stands for Repetitive Strain Injury. This disorder involves damage to muscles, tendons and nerves caused by overuse or misuse, and affects the hands, wrists, elbows, arms, shoulders, back, or neck". This answer provides the user with relevant background information about the topic of the

[1] For more information about IMOGEN, see http://wwwhome.cs.utwente.nl/~theune/IMOGEN/index.html/.

question. In addition, including additional text in the answer may allow the user to assess the answer's accuracy in order to verify whether it is correct or not (Bosma, 2005). This raises the question which kind of answer presentations (unimodal vs. multimodal) would be best for different types of questions and answers.

Much research has been done in the field of cognitive psychology on the influence of (combinations of) different modalities on the users' understanding, recall and processing efficiency of the presented material (e.g., Carney & Levin 2002, Mayer 2005, Tversky, Morrison & Betrancourt 2002). This research has resulted in several guidelines on how to present (multimodal) information to the user, such as the multimedia principle (i.e., instructions should be presented using both text and pictures, rather than text only) and the spatial contiguity principle (i.e., when presenting a combination of text + pictures, the text should be close to or embedded within the pictures) (Mayer, 2005). However, these guidelines are based on specific types of information used in specific domains in particular descriptions of cause and effect chains which explain how systems work (Mayer 1989, Mayer & Gallini 1990, Mayer & Moreno 2002) and procedural information describing how to acquire a certain skill (Marcus, Cooper & Sweller 1996, Michas & Berry 2000, Schwan & Riempp 2004). These guidelines do not tell us which modalities are most suited for which information types, as each learning domain has its own characteristics (Van Hooijdonk & Krahmer, submitted).

Several researchers have tried to make an overview of the characteristics of modalities, information types, and the matches between them. For example, Bernsen (1997) focussed on the features of modalities in his Modality Theory, i.e., *"given any particular set of information which needs to be exchanged between user and system during task performance in context, identify the input/output modalities, which, from the user's point of view, constitute an optimal solution to the representation and exchange of that information"*. He proposed a taxonomy to define generic unimodalities consisting of various features. Other researchers proposed taxonomies of information types such as dynamic, static, conceptual, concrete, spatial, and temporal in order to select the appropriate modalities (e.g., Heller, Martin, Haneef, Gievska-Krliu 2001, Sutcliffe, 1997).

Other research has been concerned with the so-called media allocation problem: *"How does a producer of a presentation determine which information to allocate to which medium, and how does a perceiver recognize the function of each part as displayed in the presentation and integrate them into a coherent whole?"* (Arens, Hovy & Vossers, 1993). According to Arens et al. (1993) the characteristics of the media used are not the only features that play a role in media allocation. The characteristics of the information to be conveyed, the goals and characteristics of the producer, and the characteristics of the perceiver and the communicative situation are also important. In order to create a multimodal information presentation, modalities should be integrated dynamically based on a communication theory as a whole (e.g., André 2000, Arens et al. 1993, Maybury & Lee 2000, Oviatt et al. 2003).

In short, several research fields have been concerned with the generation of multimodal information presentations resulting in several guidelines, frameworks, and taxonomies. However what is really needed to generate optimal multimodal presentations is gaining knowledge on whether users present information in a multimodal way, and if so, when and how they present this multimodal information. To achieve this goal, we will carry out a series of experiments following the cognitive engineering approach as used by Heiser et al. (2004). In this approach, human users are asked to produce information presentations, which are then rated by other users. Based on the results, cognitive design principles are identified and used to improve the automatic generation of information presentations. In this paper, we present a production experiment in which users' (multimodal) answers to different medical questions were collected. We expected that both question type (definition vs. procedural questions) and answer type (brief or extended) would affect the answer presentation, i.e., some answers would probably consist of text only while others would probably consist of a combination of modalities like text + picture. In a later stage, a preference experiment will be conducted in which other users will rate the answer presentations collected in the production experiment.

This paper is structured as follows. In section 2 the research method is described followed by the results of a first analysis in section 3. We end with a general conclusion in section 4.

2 METHOD

2.1 PARTICIPANTS

One hundred and eleven students of Tilburg University participated for course credits. Of the students, 46 were male and 65 were female. The mean age was 22 years (Std = 2,10). All participants used the computer for their study and had a computer at their disposal at home.

2.2 STIMULI

The participants were given one of four sets of eight general medical questions for which the answers could be found on the World Wide Web. The participants had to give two types of answers per question i.e., a brief answer and an extended answer. Besides, different (combinations of) modalities could be used to answer the questions. The participants had to assess for themselves which (combination of) modalities were best for a given question, and they were specifically asked to present the answers as they would prefer to find them in a QA system. To make sure they could carry out this task, they were instructed about the working of QA systems in advance. Questions and answers had to be presented in a fixed format in PowerPoint™ with areas for the question ("vraag") and the answer ("antwoord"). This programme was chosen because it has the possibility to insert pictures, film clips, and sound fragments in an answer presentation. All participants were familiar with PowerPoint™ and most of them used it on a monthly basis (51,4%).

Of the eight questions in each set, four were randomly chosen from one hundred medical questions formulated to test the IMIX QA system (e.g., how many X-chromosomes does a woman have in her body cells?). Of the remaining four questions, two were definition questions and two were procedural questions. Orthogonal to this, two questions referred explicitly or implicitly to body parts and two did not. These four question types were given to the participants in a random order. Examples of the questions were:

- Definition question + body parts: "Where is progesterone produced?" or "Where are red blood cells produced?"
- Definition question - body parts: "What are the side effects of ibuprofen?" or "What are thrombolytic drugs?"
- Procedural question + body parts: "How should a sling be applied to the left arm?" or "What should be done when having a nosebleed?"
- Procedural question - body parts: "What happens when a myelogram is taken?" or "How is a SPECT scan made?"

2.3 CODING SYSTEM

Each answer was coded as belonging to a category of the following variables: the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, animations, the function of these visual media[2] related to text, and the position of the visual media related to text. Our coding criteria for these variables are discussed below. To determine the reliability of the coding system, Cohen's κ (Krippendorff, 1980) was calculated. However, the Cohen's κ was not calculated for two of the variables: number of words and text. The number of words was counted automatically; therefore no agreement had to be established between the annotators. Second, text occurred in 98% of the answers. The remaining 2% of the answers were insufficient to determine the Cohen's κ. Below we will describe our criteria for coding the answers.

Number of words: The number of words was counted automatically.

Text: We distinguished the presence of textual answers (i.e., answers that contained text, possibly in combination with other media) versus non-textual answers.

[2] By visual media we mean photos, graphics, and animations.

Typographic manipulation: An answer contained typographic manipulation if the following features occurred: the use of bold, italic, underlining, or colour in the text of the answer.

Spatial manipulation: An answer contained spatial manipulation if the following features occurred: dividing the text into sections, indenting the text, using headings, or using enumeration.

Graphical manipulation: An answer contained graphical manipulation if the following features occurred: using tables, horizontal or vertical lines, arrows, or bullets.

Photos: We distinguished whether the answer contained no photo, one photo or several photos.

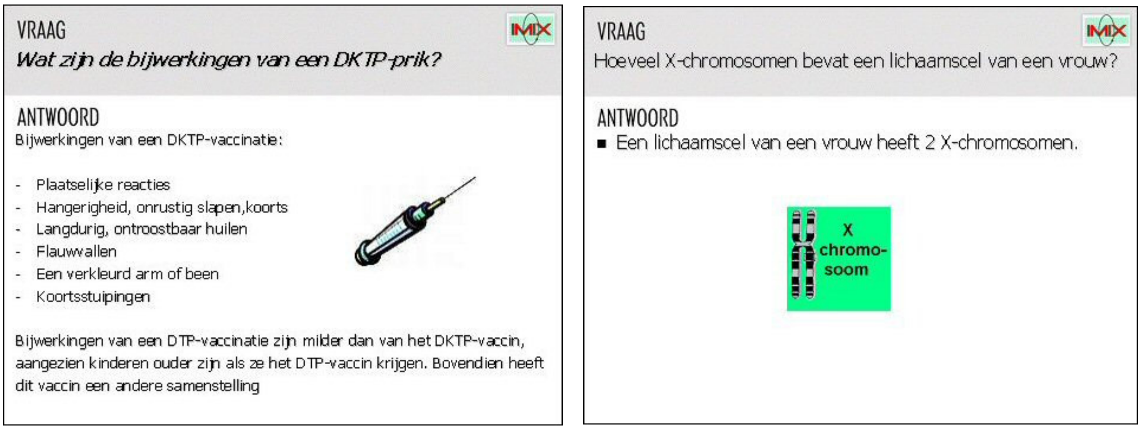
Graphics: We defined graphics as non-photographic, static depictions of concepts (e.g., diagrams, charts, and line drawings). We distinguished whether the answer contained no graphic, one graphic, or several graphics.

Animations: We defined animations as dynamic visuals possibly with sound (e.g., film clips and animated pictures). We distinguished whether the answer contained no animations, one animation, or several animations.

Position of visual media: We wanted to know what the position of the visual media (i.e., photos, graphics, and animations) was compared to the text within the answer presentations. We distinguished whether the visual media were in the upper, lower, left or right part of the answer area.

Function of visual media: We wanted to know what the function of the visual medium (i.e., photos, graphics, and animations) was in relation to text within the answer presentations. We distinguished three functions, loosely based on Carney & Levin (2002):

1. *Decorational function*: a visual medium has a decorational function if removing it from the answer presentation does not alter the informativity of the answer in any way. Figure 1 shows two examples of answer presentations in which the visual medium has a decorational function. The example on the left shows an answer to the question: "What are the side effects of a vaccination for diphtheria, whooping cough, tetanus, and polio?" Within the answer spatial manipulation occurs (i.e., the text is divided into sections and an enumeration is used). Also graphical manipulation occurs (i.e., bullets are used). The answer consists of a combination of text + graphic. The text describes the side effects of the vaccination while the graphic only shows a syringe. The graphic does not add any information to the answer; therefore it has a decorational function. The example on the right shows an answer to the question: "How many X-chromosomes are there in a woman's body cell?" The answer consists of a combination of text + graphic. In text the answer is given (i.e., a woman's body cell has two X-chromosomes). The answer would not be less informative if the graphic was not present.



VRAAG

Wat zijn de bijwerkingen van een DKTP-prik?

ANTWOORD

Bijwerkingen van een DKTP-vaccinatie:

- Plaatselijke reacties
- Hangerigheid, onrustig slapen,koorts
- Langdurig, ontroostbaar huilen
- Flauwvallen
- Een verkleurd arm of been
- Koortsstuipingen

Bijwerkingen van een DTP-vaccinatie zijn milder dan van het DKTP-vaccin, aangezien kinderen ouder zijn als ze het DTP-vaccin krijgen. Bovendien heeft dit vaccin een andere samenstelling

VRAAG

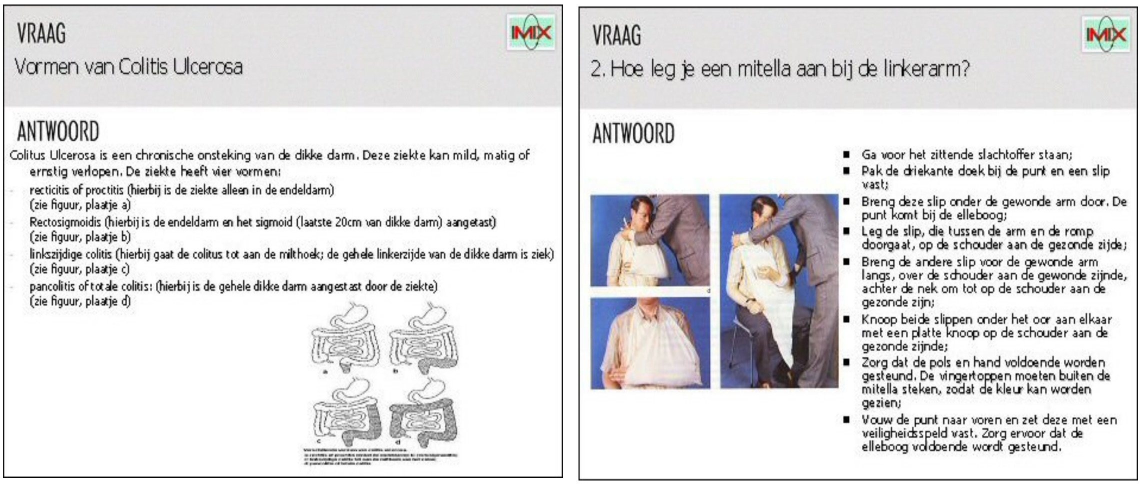
Hoeveel X-chromosomen bevat een lichaamscel van een vrouw?

ANTWOORD

■ Een lichaamscel van een vrouw heeft 2 X-chromosomen.

Figure 1: Examples of answer presentations with a visual medium having a decorational function

2. *Representational function*: a visual medium has a representational function if removing it from the answer presentation does not alter the informativity of the answer, but its presence clarifies the text. Figure 2 shows two examples of answer presentations in which the visual medium has a representational function. The example on the left shows an answer to the question: “What types of colitis can be distinguished?” Within the answer spatial manipulation (i.e., an enumeration is used) and graphical manipulation occurs (i.e., bullets are used). The answer consists of a combination of text + graphic. The text describes the four types of colitis and their occurrence in the intestines. This information is visualized in the graphics. The example on the right shows an answer to the question: “How should a sling be applied to the left arm?” The answer consists of three photos illustrating the procedure, which is described in more detail in the text on the right.



VRAAG

Vormen van Colitis Ulcerosa

ANTWOORD

Colitus Ulcerosa is een chronische onsteking van de dikke darm. Deze ziekte kan mild, matig of ernstig verlopen. De ziekte heeft vier vormen:

- recticitis of proctitis (hierbij is de ziekte alleen in de endeldarm) (zie figuur, plaatje a)
- Rectosigmoidis (hierbij is de endeldarm en het sigmoid (laatste 20cm van dikke darm) aangetast) (zie figuur, plaatje b)
- linkszijdige colitis (hierbij gaat de colitus tot aan de milthoek; de gehele linkerzijde van de dikke darm is ziek) (zie figuur, plaatje c)
- pancolitis of totale colitis: (hierbij is de gehele dikke darm aangestast door de ziekte) (zie figuur, plaatje d)

VRAAG

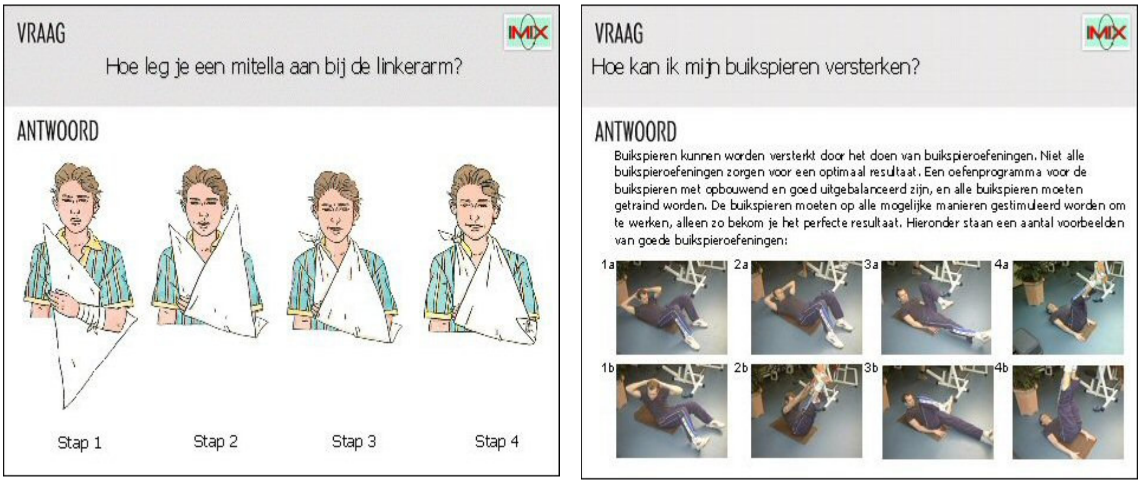
2. Hoe leg je een mitella aan bij de linkerarm?

ANTWOORD

- Ga voor het zittende slachtoffer staan;
- Pak de driekante doek bij de punt en een slip vast;
- Breng deze slip onder de gewonde arm door. De punt komt bij de elleboog;
- Leg de slip, die tussen de arm en de romp doorgaat, op de schouder aan de gezonde zijde;
- Breng de andere slip voor de gewonde arm langs, over de schouder aan de gewonde zijnde, achter de nek om tot op de schouder aan de gezonde zijn;
- Knoop beide slippen onder het oor aan elkaar met een platte knoop op de schouder aan de gezonde zijnde;
- Zorg dat de pols en hand voldoende worden gesteund. De vingertoppen moeten buiten de mitella steken, zodat de kleur kan worden gezien;
- Vouw de punt naar voren en zet deze met een veiligheidsspeld vast. Zorg ervoor dat de elleboog voldoende wordt gesteund.

Figure 2: Examples of answer presentations with a visual medium having a representational function

3. *Additional function*: a visual medium has an additional function if removing it from the answer presentation alters the informativity of the answer. If an answer consists only of a visual medium, it automatically has an additional function. Figure 3 shows two examples of answer presentations in which the visual medium has an additional function. The example on the left shows the answer to the question: “How should a sling be applied to the left arm?” The answer consists of four graphics illustrating the procedure. The example on the right shows an answer to the question: “How can I strengthen my abdominal muscles?” The text describes some general information about abdominal exercises (i.e., an exercise program should be well balanced in which all abdominal muscles must be trained). The photos represent four exercises which can be done to strengthen the abdominal muscles.



VRAAG

Hoe leg je een mitella aan bij de linkerarm?

ANTWOORD

VRAAG

Hoe kan ik mijn buikspieren versterken?

ANTWOORD

Buikspieren kunnen worden versterkt door het doen van buikspieroefeningen. Niet alle buikspieroefeningen zorgen voor een optimaal resultaat. Een oefenprogramma voor de buikspieren met opbouwend en goed uitgebalanceerd zijn, en alle buikspieren moeten getraind worden. De buikspieren moeten op alle mogelijke manieren gestimuleerd worden om te werken, alleen zo bekom je het perfecte resultaat. Hieronder staan een aantal voorbeelden van goede buikspieroefeningen:

Figure 3: Examples of answer presentations with a visual medium having an additional function

2.4 ANNOTATION PROCEDURE

In total 1776 answers were collected (111 participants × 8 questions × 2 (brief, extended) answers). However, one participant gave 15 answers, resulting in one missing value. Thus, the coded corpus consisted of 1775 answers. The coding scheme (see Section 2.3) was formulated and given to six annotators (the authors plus one other annotator). The annotation was done in two steps. First, each annotator independently coded a part of the corpus to determine the adequacy of the coding scheme. Differences between the annotators were discussed, which resulted in some adjustments of the coding system. Subsequently, every annotator independently coded the same set of 112 answers. Second, every annotator independently coded a part of the total corpus (i.e., approximately 300 answers).

To compute agreement we used Cohen's κ measure. Following standard practice, Cohen's κ scores between .81 and 1.00 signify an almost perfect agreement, between .61 and .80 signify a substantial agreement, between .41 and .60 is a moderate agreement, and between .21 and .40 is a fair agreement (Rietveld & van Hout, 1993). Table I summarizes the results. The annotators corresponded in judging the occurrence of typographic manipulation. They highly corresponded in judging the occurrence of spatial manipulation, photos, graphics, and animations. Moreover, an almost perfect agreement was reached in assigning a function to the visual media, and a substantial agreement was reached in assigning a position to the visual media[3]. However, a low agreement was reached for the occurrence of graphical manipulation. A possible explanation for this result could be that the use of graphical manipulation interfered with the use of PowerPoint™. This program presents the information point by point using bullets. It was not clear whether the participants used the bullets intentionally or unintentionally to present the information. Therefore, some analysts coded the use of bullets as an occurrence of graphical manipulation and some did not, resulting in a low kappa score for this variable.

Typographic manipulation	.74
Spatial manipulation	.89
Graphical manipulation	.41
Photo's	.81
Graphics	.83
Animations	.92
Function of visual media	.83
Position of visual media	.74

Table 1: Cohen's κ scores of agreement for typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, animations, the function and the position of visual media (n = 112 answer presentations)

3 RESULTS

3.1 DESCRIPTIVE STATISTICS

Table 2 shows the frequencies of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations in the complete corpus of coded answer presentations. Inspection of Table 2 reveals that most answer presentations contain text. Almost one in five answers contained typographic manipulation. Spatial manipulation occurred in almost half of the answer presentations and graphical manipulation occurred in one of the six answer presentations. Almost one in four answers contained one or more visual media of which graphics were most frequent and animations were least frequent. The presence of photos was between these two.

[3] The Cohen's κ for the variable "position of visual media" is based on the judgments of five annotators.

Table 3 shows the frequencies of photos and graphics related to their position.[4] The analysis of the position of visual media revealed significant effects for both photos (χ^2 (3) = 75.96, p < .001) and graphics (χ^2 (3) = 176.02, p < .001). In both cases, the medium was most often placed below the text.

Table 4 shows the frequencies of photos, graphics, and animations related to their function. Note that the answer presentations in which photos, graphics, or animations co-occurred are not shown in the table. Table 3 reveals that the distribution of photos related to their function differed significantly from chance (χ^2 (2) = 42.84, p < .001). Most photos had a representational function. Also, the distribution of graphics related to their function differed significantly from chance (χ^2 (2) = 34.50, p < .001). Most graphics had a representational function. Finally, the distribution of animations related to their function differed significantly from chance (χ^2 (2) = 63.88, p < .001). Most animations had an additional function.

Text		98.3
Typographic manipulation		18.1
Spatial manipulation		47.6
Graphical manipulation		16.7
Visual media[5]		24.0
	Photos	9.0
	Graphics	14.2
	Animations	3.6

Table 2: Frequencies of text, typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, and animations in 1775 coded answers from 111 participants (Scores are percentages of answers; n = 1775)

	Position of visual media with respect to text				
	Above text	Below text	Left of text	Right of text	Totals
Photo (n = 114)	8.8	56.1	3.5	31.6	100.0
Graphic (n =193)	3.6	61.1	2.1	33.2	100.0

Table 3: Frequencies of photos and graphics related to their position (Scores are percentages of answers)

[4] The position of animations was not taken into account because they were always added to the answer presentations with hyperlinks. Moreover, in this analysis answer presentations in which only a visual medium occurred are not taken into account.

[5] In some answers several visual media occurred (i.e., photos, graphics, and animations). These instances were counted as one occurrence of visual media. Thus the sum of the frequencies of photos, graphics, and animations does not correspond with the overall frequency of the variable visual media.

	Function of visual media			
	Decorational function	Representational function	Additional function	Totals
Photos (n = 129)	20.9	60.5	18.6	100.0
Graphics (n = 221)	15.4	46.6	38.0	100.0
Animations (n = 48)	2.1	10.4	87.5	100.0

Table 4: Frequencies of photos, graphics, and animations related to their function (Scores are percentages of answers)

3.2 BRIEF AND EXTENDED ANSWERS

As expected the type of answer (brief vs. extended) affected the answer presentation. The type of answer had a significant effect on the mean number of words used (t (1726) = 30.39, p < .001). The mean number of words used in brief answers was 24 words (Std = 23,02), while the mean number of words used in extended answers was 106 words (Std = 76,20). Table 5 shows the frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations within the brief and extended answers. The results showed that there was a significant difference in the presence of all the variables within the answer type. Inspection of Table 5 reveals that they all occurred more frequently within the extended answers.

Table 6 shows the frequencies and χ^2 statistics of the functions of visual media related to brief and extended answers. The results showed that the overall distribution of the functions of visual media within the answer type differed significantly (χ^2 (2) = 31.47, p < .001). Visual media with a decorational function occurred significantly more often in brief answers than in extended answers. Visual media having a representational function occurred significantly more often in extended answers. Finally, visual media having an additional function occurred significantly more often in brief answers.

	Brief answers (n = 888)	Extended answers (n = 887)	χ^2 statistics
Text	97.5	99.0	χ^2 (1) = 5.53, p < .025
Typographic manipulation	9.8	26.5	χ^2 (1) = 83.30, p < .001
Spatial manipulation	23.9	71.4	χ^2 (1) = 401.24, p < .001
Graphical manipulation	8.9	24.5	χ^2 (1) = 77.40, p < .001
Visual media	11.0	38.0	χ^2 (1) = 174.30, p < .001
Photos	4.8	13.1	χ^2 (1) = 36.90, p < .001
Graphics	5.5	22.9	χ^2 (1) = 109.98, p < .001
Animations	.9	6.3	χ^2 (1) = 37.40, p < .001

Table 5: Frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations related to the brief and extended answers (Scores are percentages of answers; n = 1775)

	Brief answers (n = 98)	Extended answers (n = 338)	χ^2 statistics
Decorational function	25.5	13.3	χ^2 (1) = 5.71, p < .025
Representational function	21.4	53.3	χ^2 (1) = 125.78, p < .001
Additional function	53.1	33.4	χ^2 (1) = 22.55, p < .001
Totals	100.0	100.0	

Table 6: Frequencies of the function of visual media related to brief and extended answers (Scores are percentages of answers; n = 436)

3.3 DEFINITION AND PROCEDURAL QUESTIONS WITH AND WITHOUT REFERENCE TO BODY PARTS

We were interested whether different types of questions were related to different answer presentations. Therefore we analyzed a subset of the medical questions (i.e., the definition and procedural questions with and without reference to body parts). The results indicated that the type of question had a significant effect on the mean number of words used in the answer presentation (F (3, 484.63) = 9.28, p < .001)[6]. Post hoc tests indicated that the answers of procedural questions consisted of more words than the answers of definition questions irrespective of reference to body parts.

Table 7 shows the frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations within the definition and procedural questions and within questions with and without reference to body parts. The results showed that the distribution of typographic manipulation within the question types did not differ: typographic manipulation occurred equally within all question types. However, the distribution of all other variables within the question types differed significantly. Text occurred most frequently within definition questions with reference to body parts and procedural questions without reference to body parts. Spatial and graphical manipulation were most frequent in procedural questions with reference to body parts. The use of visual media was also most frequent within this type of questions. Finally, photos and animations occurred more often in answers to procedural questions with reference to body parts. However, graphics occurred more often in answers to *definition* questions with reference to body parts.

Table 8 shows the frequencies and χ^2 statistics of the functions of visual media within definition and procedural questions and within questions with and without reference to body parts. The results show that the functions of visual media differed significantly within the question types (χ^2 (6) = 91.84, p < .001). Table 8 shows that visual media with a decorational function occurred most often in definition questions *without* reference to body parts, and that visual media with a representational function occurred most often in definition questions *with* reference to body parts. Finally, visual media having an additional function occurred most often in *procedural* questions with reference to body parts.

4 CONCLUSION

In this paper we described a production experiment following a cognitive engineering approach in order to gain knowledge on which modality combinations are used in manually created answers. A total of 1775 answers to different medical questions were collected. These answers were coded as belonging to a category of the following variables: text, typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, animations, the function and the position of the visual media related to text. To determine the reliability of this coding scheme, six annotators coded part of the data. The results of this reliability analysis indicated that for most variables the annotators corresponded highly in their judgments.

[6] Tests for significance were performed using a Welch's analysis of variance (ANOVA), as the variances were not equal. A significance threshold of .05 was used and for post hoc tests the Tukey HSD method was used.

	Definition questions (n = 443)		Procedural questions (n = 444)		
	Body parts (n = 222)	¬ Body parts (n = 221)	Body parts (n = 222)	¬ Body parts (n = 222)	χ^2 statistics
Text	99.5	99.1	94.1	99.5	χ^2 (3) = 24.61, p < .001
Typographic manipulation	22.5	16.3	18.0	17.1	χ^2 (3) = 3.421, p = .33
Spatial manipulation	38.7	53.4	61.3	41.4	χ^2 (3) = 29.47, p < .001
Graphical manipulation	9.5	20.8	29.7	9.5	χ^2 (3) = 44.83, p < .001
Visual Media	31.1	10.4	46.8	32.4	χ^2 (3) = 70.84, p < .001
Photos	4.5	5.9	24.3	19.8	χ^2 (3) = 55.73, p < .001
Graphics	28.4	5.0	13.1	11.7	χ^2 (3) = 52.29, p < .001
Animations	.5	.9	13.5	5.0	χ^2 (3) = 51.74, p < .001

Table 7: Frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations within the definition and procedural questions and within questions with and without reference to body parts.

	Definition questions (n = 92)		Procedural questions (n = 177)		
	Body parts (n = 69)	¬ Body parts (n = 23)	Body parts (n = 105)	¬ Body parts (n = 72)	χ^2 statistics
Decorational function	5.8	65.2	4.8	5.6	χ^2 (3) = 12.29, p < .01
Representational function	63.8	21.8	41.0	54.2	χ^2 (3) = 31.78, p < .001
Additional function	30.4	13.0	54.2	40.2	χ^2 (3) = 55.09, p < .001
Totals	100.0	100.0	100.0	100.0	

Table 8: Frequencies and χ^2 statistics of the functions of visual media related to the definition and procedural questions and to questions with and without reference to body parts (Scores are percentages of answers; n = 269)

A first analysis of the data showed that the participants used combinations of text and visual media to present their answers. Almost one in four answers contained one or more visual media. Moreover, significant differences were found in the distribution of photos, graphics, and animations related to their function. Photos often had a representational function: they visually represented the information mentioned in the text. Animations often had an additional function because they present the information dynamically as opposed to photos. Graphics often had either a representational or an additional function. A possible explanation for this result could be that graphics are more diverse. While some graphics visually represent the information mentioned in text, other graphics represent information in such a way (e.g., the presence of arrows or charts) that they contain more information than mentioned in the text.

As expected the type of answer (brief vs. extended) affected the answer presentation. Extended answers consisted of more words than the brief answers, but also word manipulation, spatial manipulation and graphical manipulation were more frequent in the extended answers. A possible explanation for this result could be that presenting more text affects the readability. Typographic manipulation, spatial manipulation, and graphical manipulation could help to make the text more transparent and thus more readable. Also visual media were more frequent in the extended answers. Within brief answers, most frequent were visual media with a decorational and additional function whereas visual media with a representational function were more frequent within extended answers. A possible explanation for this result could be that when the answer does not contain much text, it is likely that the visual medium will have an additional function (i.e., it expresses more information). When the answer contains much text, it is likely that the visual medium will have a representational function (i.e., it represents the information mentioned visually).

The type of question also affected the answer presentation. Answers to procedural questions consisted of more words. Besides, spatial and graphical manipulation occurred more frequently in answers to this type of questions. A possible explanation for this result could be that procedural information consists of several steps that have to be described. Moreover, dividing the text into sections or using headings may help the user to see when one step ends and another begins (Ganier, 2004). The distribution of visual media differed significantly within the question types. Photos and animations occurred most often in answers to procedural questions with reference to body parts. These visual media may help to visualise the steps of a procedure. However, graphics occurred most often in answers to definition questions with reference to body parts. As mentioned earlier, graphics are more diverse making them perhaps more suitable for other question types. For example, the definition question "Where is testosterone produced?" may be more clearly visualized with a graphic in which different parts of the male reproductive system are illustrated.

The first results of this production experiment following the cognitive engineering approach showed that users do make use of multiple media in their information presentations and that the design of these presentations is affected by the answer and question type. However what is not clear is which kind of multimodal information presentation users would prefer. This will be tested in a second experiment in which users will rate the answer presentations collected in this production experiment. Based on the results of this experiment, we intend to develop a set of design principles for multimodal answer presentation in the IMIX medical QA system.

ACKNOWLEDGEMENTS

The current research is performed at the Communication and Cognition group at Tilburg University and the Human Media Interaction group of the University of Twente within the IMOGEN (Interactive Multimodal Output GENeration) project. IMOGEN is a project within the Netherlands Organisation for Scientific Research (NWO) research programme on Interactive Multimodal Information eXtraction (IMIX). The authors would like to thank Paul Flapper for his help in setting up the experiment, Jurry de Vos for his help in annotating the data, and Jeroen Geertzen for his help with the statistics.

REFERENCES

André, E. (2000). The generation of multimedia presentations. In R. Dale, H. Moisl, & H. Somers (Eds.) *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker Inc., NY. pp. 305-327.

Arens, Y., Hovy, E. & Vossers, M. (1993). On the knowledge underlying multimedia presentations. In M. T. Maybury (Ed.) *Intelligent Multimedia Interfaces*, AAAI Press, Menlo Park, CA, pp. 280 - 306.

Bernsen, N. (1994). Foundations of multimodal representations. A taxonomy of representational modalities. *Interacting with Computers 6* (4), pp. 347-371.

Bosma, W. (2005). Extending answers using discourse structure. In *Proceedings of the RANLP Workshop on Crossing Barriers in Text Summarization Research*. H. Saggion & J.-L. Minel (Eds), Incoma Ltd., Borovets, Bulgaria, pp. 2-9.

Carney, R. & Levin, J. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review 14*(1), pp. 5-26

Ganier, F. (2004). Factors affecting the processing of procedural instructions: implications for document design. *IEEE Transactions on Professional Communication. 47* (1), pp. 15 – 26.

Heiser, J., Phan, D., Agrawala, D., Tversky, B. & Hanrahan, P. (2004). Identification and validation of cognitive design principles for automated generation of assembly instructions. In *Proceedings of the Working Conference on Advance Visual Interfaces*, ACM Press, NY, pp. 311-319.

Heller, R., Martin, C., Haneef, N. & Gievka-Krliu, S. (2001). Using a theoretical multimedia taxonomy framework. *ACM Journal of Educational Resources in Computing 1* (1), pp. 1-22.

Hooijdonk, C.M.J. van & Krahmer. E. (submitted). Information modalities for procedural instructions: the influence of text, static and dynamic visuals on learning and executing RSI exercises.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, USA.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology, 88* (1), pp. 49-62.

Maybury, M. & Lee, J. (2000). Multimedia and multimodal interaction structure. In M. Taylor, F. Néel & D. Bouwhuis (Eds.). *The Structure of Multimodal Dialogue II*, John Benjamins, Amsterdam, pp. 295-308.

Mayer, R. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology 81* (2), pp. 240-246

Mayer, R. & Gallini,. J. (1990). When is an illustration worth a thousand words? *Journal of Educational Psychology, 82,* pp. 715-726,

Mayer, R. & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and Instruction, 12*, 107-119.

Mayer, R. E. (2005). *The Cambridge Handbook of Multimedia Learning*. Cambridge [etc.]: Cambridge University Press.

Michas, I. & Berry, D. (2000). Learning a procedural task: effectiveness of multimedia presentations, *Applied Cognitive Psychology, 14*, pp. 555-575.

Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M. & Carmichael, L. (2003). Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, Vancouver, British Columbia, Canada, pp. 44–51.

Rietveld, T. & Hout, R. van (1993). *Statistical Techniques for the Study of Language and Language Behavior*. Berlin: Mouton de Gruyter.

Sutcliffe, A. (1997). Task-related information analysis. *Int. Journal of Human Computer Studies, 47* (2), 223-257.

Schwan, S. & Riempp, R. (2004). The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning and Instruction, 14*, pp. 293-305.

Theune, M., Schooten, B. van, Akker, R. op den, Bosma, W., Hofs, D., Nijholt, A., Krahmer, E., Hooijdonk, C. van & Marsi, E. (to appear). Questions, pictures, answers: introducing pictures in question-answering systems. To appear in *Proceedings of the Tenth International Symposium on Social Communication*, 22-26 January 2007, Santiago de Cuba, Cuba.

Tversky, B., Morrison, J. & Betrancourt, M. (2002). Animation: can it facilitate? *Int. Journal of Human Computer Studies, 57*, 247-262.

Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System*

Erwin Marsi, Ferdi van Rooden
Communication and Cognition
Tilburg University
The Netherlands
e.c.marsi@uvt.nl

Abstract

One of the strategies that question-answering (QA) systems may follow to retain users' trust is to express the level of uncertainty attached to answers they provide. Multimodal QA systems offer the opportunity to express this uncertainty through other than linguistic means. On the basis of evidence from the literature, it is argued that uncertainty is in fact better expressed by audiovisual than by verbal means. We summarize unpublished work on audiovisual expression of uncertainty in the context of QA systems which suggests that users prefer visual over linguistic signaling. Next, we describe a perception experiment showing that uncertainty can be reliably expressed by means of a talking head using a limited repertoire of animated facial expressions, i.e. only combinations of eyebrow and head movements. In addition, we discuss a number of open issues that need to be resolved before a talking head can really be employed for signaling uncertainty in multimodal human-computer interaction.

Keywords: certainty, confidence, trust, facial expression, facial animation, embodied conversational agents, talking heads, multimodal dialogue, question answering

1 Introduction

A commonly held opinion among researchers in the field on automatic question answering (QA) is that "incorrect answers are worse than no answers" (Burger et al., 2003). Incorrect answers evidently make the system look unreliable and undermine the user's trust in its capabilities. Since flawless QA systems are unlikely to appear soon, strategies are required to retain the user's trust. Recent QA tracks in the TREC evaluations have included questions that have no answers in the underlying data collection, forcing systems to 'know' that they are not certain of an answer (Voorhees, 2003). Other approaches include providing additional context so users can make their own judgments regarding the reliability of the answer's source (Lin et al., 2003), associating *trust values* to source documents and using these to calculate trust values for answers based on them (Zaihrayeu et al., 2005), or explaining how the answer was derived (Moldovan et al., 2003).

In this work, we explore yet another aspect of coping with uncertainty in QA systems (as a matter of fact, none of the approaches mentioned are mutually exclusive). It was carried out in the context of the IMIX project, which aims at building a multimodal QA system capable of answering questions in the medical domain, especially about Repetitive Strain Injury (RSI) (Boves and den Os, 2005; Theune et al., 2007). The IMIX demonstrator produces multimodal output in the form of text and pictures, as well as speech output and facial animation. The latter relies on the Nextens speech synthesizer for Dutch in cooperation with the RUTH talking head (DeCarlo

[*] This work was carried out within the IMIX-IMOGEN (Interactive Multimodal Output Generation) project, sponsored by the Netherlands Organization of Scientific Research (NWO). The authors like to thank Emiel Krahmer, Marc Swerts and colleagues for kindly permitting us to summarize their unpublished work on the preferred modality to express uncertainty in Section 2.2. We also like to acknowledge the useful comments and suggestions of three anonymous reviewers on our initial abstract.

and Stone, 2003; DeCarlo et al., 2004). The system incorporates multiple QA engines, some of which are capable of attaching confidence levels to their answers, albeit not always reliably. We are interested in the best way to convey uncertainty in the context of such a multimodal QA system, which offers the opportunity to exploit other communication channels besides text. In particular, the question addressed in this work is whether we can express uncertainty by means of talking head.

The remainder of the paper is organized as follows. In Section 2, we elaborate on the background and context of this work. We argue – on the basis of evidence from human-human dialogue studies – that uncertainty is better expressed by visual means than by text only. We summarize an unpublished study on audiovisual expression of uncertainty in the context of QA systems. We also discuss related work on *trust*. Section 3 reports on an experiment to test whether we can reliably express certainty or uncertainty by means of a limited repertoire of animated facial expressions, in particular, only combinations of eyebrow movements and head movements were considered. The results are in principle positive, but a number of remaining problems are discussed. In the final Section we summarize our findings and finish with a general discussion of open issues that need to be addressed before we can actually apply this approach in a multimodal QA system.

2 Background

2.1 Uncertainty in Human-Human Dialogue

In human-human information seeking dialogue, the information exchange is usually not limited to facts, but includes all sorts of additional meta-information. This kind of meta-information is often expressed by non-verbal means such as speech prosody, facial expression or gesture (e.g. Burgoon, 1994). One important example of this is the level of confidence or certainty associated with a particular piece of information. A number of researchers have used the *Feeling of Knowing* (FOK) paradigm (Hart, 1965) to study production and perception of uncertainty in human question answering. Smith and Clark (1993) found that speakers signal uncertainty regarding the correctness of their answer by means of prosodic cues such as filled pauses, increased delays and rising intonation. Subsequently, Brennan and Williams (1995) showed that listeners use these prosodic cues to estimate the level of certainty of a speaker's answer, suggesting that Smith and Clark's uncertainty cues do indeed have communicative relevance. Recent work by Swerts, Krahmer and colleagues has extended this line of research to audio-visual prosody, and in particular facial cues to uncertainty (Swerts et al., 2003; Krahmer and Swerts, 2005; Swerts and Krahmer, 2005). They found that in addition to the auditory cues, there are a number of facial cues that speakers produce to signal their uncertainty about an answer, and that those same signals are perceived by listeners in order to reliably detect the level of certainty associated with answers. Furthermore, detecting uncertainty turned out to be easier with bimodal presentation (i.e. both speech and face) in comparison with unimodal presentation. It is suggested that these findings have potential for improving human-computer interaction.

2.2 Preference for Visual versus Linguistic Cues

In recent, hitherto unpublished work, Krahmer et al. studied the expression of uncertainty in the context of a QA system. Since to the best of our knowledge no other published work addresses this topic, we will summarize their work here. The main questions were whether users appreciate it at all when a QA system signals its level of confidence regarding the answer, and whether users prefer signaling by either linguistic or visual means. In an experiment subjects were shown screenshots of a fancy-looking – but non-existent – medical QA system ("MediQuest TM"), each one containing both a question and an answer. The questions (e.g., "What is anesthesia?") were intentionally not that hard, so subjects were expected to recognize correct answers ("The process of blocking the perception of pain and other sensations."). Of the 20 answers presented, 13 were in fact correct and 7 were incorrect. The 75 subjects were equally divided in three groups, the first of which received no signaling of uncertainty at all, the second received signaling by linguistic cues, and the third by visual cues. Signaling uncertainty by linguistic cues comprised the use of modal expressions (e.g. “I think it is the process of blocking the perception of pain and other sensations.”). For visual signaling of uncertainty, the equivalent of a thermometer was used to express the degree of certainty. The majority of the correct answers (11 out of 13) were signaled as *certain*, whereas the majority of the incorrect answers (5 out of 7) were signaled as *uncertain.*

Subjects were asked to judge

1. the *formulation* of the answer,
2. the *adequacy* of confidence signaling, and
3. overall *quality* of the answer

on a 7-point scale. The results showed that answers containing linguistic signaling of uncertainty scored significantly worse on *formulation* than their certain counterparts. No such effect was found in case of visual signaling of uncertainty. The ratings on *adequacy* showed a strong negative effect in case of an inconsistent visual signal, i.e. thermometer indicating low confidence for a correct answer or thermometer indicating high confidence for an incorrect answer. This negative effect was much smaller in case of inconsistent linguistic cues. The overall *quality* scores also showed that answers with linguistic cues for uncertainty were judged significantly worse than their counterparts with visual signaling of uncertainty.

Although the choice of domain in combination with the sometimes less subtle linguistic expression of uncertainty might have affected the results to a certain extent, a likely interpretation of these findings is that subjects disliked linguistic signaling of uncertainty and preferred visual cues instead.

2.3 Trust

A QA system that is able to indicate confidence levels for its answers is arguably perceived as more trustworthy than a system lacking this capability. In that sense, expressing uncertainty is related to trust. Interestingly, several studies suggest that audiovisual communication enhances trust in comparison with text-only communication. Riegelsberger et al. (2005) showed that humans tend to have a media bias towards audio and video advice rather than text-only advice while seeking expert advice. Work by Cassell et al, (e.g. Cassell and Bickmore, 2000)), points out that believable Embodied Conversational Agents are an important factor in building trust relations between humans and computers.

3 Experiment

3.1 Design

The goal of the experiment was to test whether we can reliably express certainty or uncertainty by means of a limited repertoire of animated facial expressions. Only combinations of eyebrow movements and head movements were considered. The experiment was designed to test three hypotheses:

1. humans notice a difference between certain and uncertain animated facial expressions;
2. humans correctly recognize animated facial expressions as certain or uncertain;
3. humans are more sensitive to eyebrow movements than to head movements as a cue for certainty.

The first hypothesis states that the difference between animations intended as certain or uncertain is at least perceivable, whereas the second hypothesis states that certain and uncertain animations are recognized as intended.

Animations with either certain or uncertain facial expressions were produced by means of three different combinations of cues: (1) primarily eyebrow movements; (2) primarily head movements;

(3) both eyebrow and head movements. This amounts to six different conditions. To minimize the effect of semantics and prosody, these conditions were tested with ten different sentences.

Animations were presented to human judges with the question *How certain do you think the speaker is of the provided answer?*. Judgments were recorded on a 5 point scale, ranging from *uncertain* (1) to *certain* (5).

3.2 Material

The text material consisted of ten question-answer pairs from the domain of Repetitive Strain Injury (RSI); see Table 3 for two examples. This choice was motivated by the desire to apply our findings in future versions of the IMIX demonstrator system. The questions were taken from the list of target questions occurring in the functional specifications of the first version of the IMIX system. The answers are full sentences which were manually extracted from the shared text material as available to the IMIX QA systems. Answers are always correct, but in some cases the formulation is suboptimal given that the original context is removed. The answers are nevertheless typical for real output of a multimodal QA system.

As our talking head, we used RUTH (Rutgers University Talking Head), a freely available cross-platform real-time facial animation system (DeCarlo and Stone, 2003; DeCarlo et al., 2004). RUTH allows one to markup text with synchronized annotations for intonation and facial movements, including eyebrow and head movements, eye blinks and smiles. It relies on the Festival text-to-speech system (Black et al., 2002) to produce the speech. We ported RUTH to Dutch, using the Festival-based Nextens TTS system to produce Dutch speech.[1]

Answers were first annotated for intonation. The original English version of RUTH relies on the ToBI (Tone and Break Indices) system, the de facto standard for annotating American-English intonation. However, as Dutch intonation is significantly different, we used the equivalent system for annotating Dutch intonation (Gussenhoven, 2005), known as ToDI (Transcription of Dutch Intonation), which is supported by the Nextens TTS system for Dutch. Two examples are given in Table 3. One of the main differences is that there is no notion of *phrasal tone* or *intermediary phrase* in ToDI; there are only pitch accents and intonational phrases. Suitable locations for pitch accents and intonational phrase boundaries were determined by the first author (who has significant experience with annotation and prediction of Dutch intonation). Non-final intonational phrases start with a low initial boundary tone (%L) and end in a high final boundary tone (H%), whereas final phrases also end in a low tone (L%). All pitch accents are realized as H*L; subsequent pitch accents within an intonational phrase are downstepped (!H*L). This annotation results in arguably the most default and unmarked pitch contour in Dutch.

Next, answers were annotated for facial expressions, which in our case was limited to the commands for eyebrow and head movements as presented in Table 1. These movements come in two types. *Batons* highlight a single word and are indicated by a final star symbol. For example, *4** signals a frown associated with a single word. *Underliners* accompany several successive words. Following the convention for intonational phrases, we use an initial and final percent symbol to signal the start and end of an underliner respectively. For instance, *%4* followed by *4%* signals a frown stretching over several words; cf. the examples in Table 3. Figure 1 provides some illustrations of RUTH head movements. For details on how these abstract specifications are realized as facial expressions in RUTH, see DeCarlo and Stone (2003).

In order to create (un)certain animations we adhered to the guidelines in Table 2 as derived from the literature (Chovil, 1991a,b; Poggi, 2002; McClave, 2000; Swerts et al., 2003; Krahmer and Swerts, 2005; Swerts and Krahmer, 2005). The notion of *new information* was in practice considered as information not previously mentioned in the question. Evidently, there is a substantial gap between these global trends and the detailed specifications required by RUTH, in particular with respect the number and alignment of movements. Our annotations are therefore to a certain extent the result of what looked right and natural to the authors within the limits of the above guidelines.

[1] `http://nextens.uvt.nl`

Value:	Effect:
1+2	raises brows
4	frowns
D	nods downward
U	nods upward
F	brings the whole head forward
B	brings the whole head backwards
L	turns to model's left
R	turns to model's right
J	tilts the whole head clockwise
C	tilts the whole head counterclockwise
DR	nods downward with some rightward motion
UR	nods upwards with some rightward motion
DL	nods downward with some leftward motion
UL	nods upwards with some leftward motion
TL	tilts clockwise with downward nodding
CL	tilts counterclockwise with downward nodding

Table 1: RUTH commands for controlling eyebrow and head movements

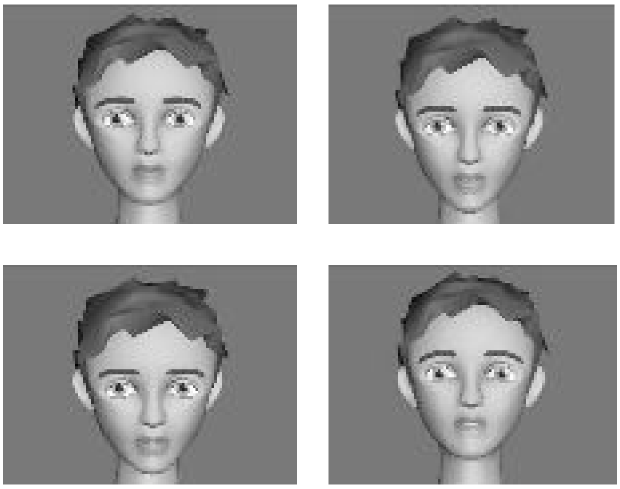


Figure 1: Illustration of several of RUTH's head movements: neutral (top left), downward nod (top right), downward nod with some leftward movement (bottom left), and upward nod (bottom left)

	Eyebrows:	Head:
Certain:	– few movements – frown with new information	– few movements – nodding with new information
Uncertain:	– many (unnecessary) movements – raising eyebrows with new information	– many (unnecessary) movements – sideward movement (shaking) with new information

Table 2: Guidelines for expressing (un)certainty through eyebrow and head movement

A related issue is that we found that animations lacking any eyebrow or head movements are almost as strange and artificial as animations without lip and jaw movements. We therefore avoided creating animations with only eyebrow movements or only head movements. Instead, all animations have at least some eyebrow and head movements, roughly corresponding to what are called *conversational facial signals* in DeCarlo et al. (2004). We used the following rules of thumb:

- Movements frequently occur with focused information – which is accented as well – and less frequently with unfocused information – which is unlikely to carry pitch accent.
- Syntactic connectives (e.g. *and, or, because*) may trigger movement, in particular when they are contrastive (e.g. *however, but, on the other hand*).
- Elements of a list may be indicated by sideward movement of the head, alternating leftward and rightward movements.
- Punctuation symbols like comma's and colons are often accompanied by a slow movement; periods often trigger a frown and/or nod; questions marks are associated with upward movement of the head and raising of the eyebrows.

The resulting RUTH animations were checked by the authors. Animations that were for some reason unnatural (e.g. suboptimal synchronization between speech and movements) were adapted. Pronunciation errors were fixed by adding words to the user lexicon.

Finally, the animations were saved as sequences of TIFF image files. The aligned synthetic speech was saved as an audio file and converted to MP3 format. Next, Adobe Premiere video editing software was used to convert images and sound to an AVI movie compressed with a standard MS Windows codec.

Question 1:	Words:	Wat	is	RSI?								
	Gloss:	what	is	RSI								
Answer 1:	Words:	RSI	is	een	**beroepsziekte**	**bij**	**mensen**	**die**	**steeds**	**dezelfde**	**beweging**	**uitvoeren**
	Gloss:	RSI	is	a	professional-disorder	of	people	who	repeatedly	the-same	movement	perform
Intonation:	Accents:				H*L					!H*L	!H*L	
	Boundaries:	%L										L%
Brows uncertain:	Brows:	1+2*			1+2*				1+2*	%1+2		1+2%
	Head:	DR*		%DL	DL%		U*		J*		%U	U%
Brows certain:	Brows:			%4	4%				4*		%4	4%
	Head:	DR*		%DL	DL%		U*		J*		%U	U%
Head uncertain:	Brows:				1+2*					%1+2		1+2%
	Head:	TL*			%TR		TR%		%L	L%		R*
Head certain:	Brows:				1+2*					%1+2		1+2%
	Head:	B*		%DL	DL%		D*		%DR			DR%
Brows & head uncertain:	Brows:	1+2*			1+2*				1+2*	%1+2		1+2%
	Head:	TL*			%TR		TR%		%L	L%		R*
Brows & head certain:	Brows:			%4	4%				4*		%4	4%
	Head:	B*		%DL	DL%		D*		%DR			DR%

Question 2:	Words:	Wie	kan	RSI	krijgen?							
	Gloss:	who	can	RSI	get							
Answer 2:	Words:	Het	is	bekend	dat	RSI	**vaker**	**voorkomt**	**bij**	**vrouwen**	**en**	**jongeren**
	Gloss:	it	is	known	that	RSI	more-often	occurs	with	women	and	youths
Intonation:	Accents:			H*L			H*L			!H*L		!H*L
	Boundaries:	%L		H%	%L							L%
Brows uncertain:	Brows:			1+2*			%1+2	1+2%		1+2*		1+2*
	Head:	DR*			D*		%TL	TL%			TR*	
Brows certain:	Brows:						%4	4%		%4		4%
	Head:	DR*			D*		%TL	TL%			TR*	
Head uncertain:	Brows:			%4	4%		%1+2	1+2%			1+2*	
	Head:	DR*					U*			L*		R*
Head certain:	Brows:			%4	4%		%1+2	1+2%			1+2*	
	Head:			D*			U*			%D		D%
Brows & head uncertain:	Brows:			1+2*			%1+2	1+2%		1+2*		1+2*
	Head:	DR*					U*			L*		R*
Brows & head certain:	Brows:						%4	4%		%4		4%
	Head			D*			U*			%D		D%

Table 3: Specifications for two of the stimuli (see text for explanation)

3.3 Procedure

A pilot experiment made clear that presenting 60 animations to a single subject takes too much time and is not feasible. The material was therefore split into six different parts, each part presenting all conditions for two different sentences.

The experiment was presented as a sequence of web pages and ran through the internet, allowing subjects to use a standard computer with a broadband internet connection and a current web browser. Subjects were automatically assigned to one of the six parts of the experiment. The introduction page explained the purpose and procedure of the experiment, and asked for some personal information (gender, age, etc.). Another page played a test animation to check that sound and image were correctly received.

Next, the stimuli were presented in random order, each one on a separate web page containing four elements. At the top of the page, there was an embedded movie player for rendering the RUTH animation. Subjects could replay this animation as many times a they liked. Below it was a plain text version of the answer to make sure that subjects understood the answer, even in case the speech synthesis was imperfect. We decided not to show the original question to prevent subjects from focusing to much on the factual content instead of on the visual presentation. At the bottom of the page was a 5 point scale (in the form of radio buttons), ranging from *sure* to *unsure*, through which subjects could respond to the question *How certain do you think the speaker is of the provided answer?.* Finally, there was a button for going to the next page. Returning to previous pages was impossible.

The closing pages offered space to provide general comments, and thanked subjects for their time.

3.4 Subjects

The online experiment was visited by 77 people, of which 58 completed a valid run. To keep the number of participants per part evenly balanced, only 50 results were used for analysis. Subjects' age ranged from 20 to 70 years old ($x = 30.6$, $SD = 11.0$); 31 were male and 19 were female. All were native speakers of Dutch without hearing impairments.

3.5 Results

The results are summarized in Table 4. Testing deviation from the expected mean (middle of the scale, i.e. 3) with a one-tailed t-test revealed that the average score on certain animations (3.63) is significantly different from the expected mean score ($p < 0.001$). This is not the case for the average score on the uncertain animations (2.85). The difference between the two scores (0.78) is again significant ($p < 0.001$). These findings confirm that overall the difference between certain and uncertain animations is at least noticeable, and that overall certain animations are recognized as intended.

Looking at nonverbal cues, we can observe that both eyebrow and head movements on their own, as well as the combination of the two, are sufficient to signal certainty (all $p < 0.001$). As far as uncertainty is concerned, however, only head movements ($p < 0.025$) and combined movements ($p < 0.01$) are close to significance. The effect of eyebrow movements is in fact opposite to the one intended. That is, eyebrow movements intended to signal *uncertainty* are actually perceived as signaling *certainty.* Thus contrary to our initial hypothesis, humans appear to be more sensitive to head movements than to eyebrow movements as far as the perception of uncertainty is concerned.

3.6 Discussion

Given the often subtle differences between the stimuli, we did not expect the differences to be significant (if noticed at all), so we think this is a rather promising result. Still, there are several issues that deserve discussion.

To begin with, we can think of alternative explanations of these results. One simple hypothesis is that more movement is perceived as more less certain, and conversely, less movement as more certain. This is not directly compatible with our results however. If we compare the total number

Cue:	Certain				Uncertain:			
	n:	av:	SD:	$p <$:	n:	av:	SD:	$p <$:
Eyebrow movements	10	3.49	0.73	.0001	10	3.26	0.82	.05
Head movements	10	3.54	0.81	.0001	10	2.65	0.95	.025
Eyebrow & head movements	10	3.91	0.77	.0001	10	2.62	0.94	.01
Overall:	30	3.63	0.58	.0001	30	2.85	0.64	n.s.

Table 4: Average scores of perceived certainty on a five point scale (uncertain=1, certain=5) over all subjects (N=50), split according to non-verbal cues used and animation's intended meaning (certain vs. uncertain); p-scores indicate significant difference from the expected mean score (3) according to a one-tailed t-test

of head movements – both batons and underliners – in the *uncertain* animations (55) to the total number of head movements in the *certain* animations (43), the difference is relatively small (12), but nevertheless sufficient to be perceived as significantly different. In contrast, the difference between the total number of eyebrow movements in *uncertain* animations (46) versus in *certain* animations (29) is slightly larger (17), yet insufficient to cause a similar significant difference in perception.

Perhaps then eyebrow movements are irrelevant for expressing uncertainty, and the results depend solely on head movements. This would explain the outlier in the case of uncertainty expressed by eyebrow movements, and is also compatible with the fact that there is hardly any difference between uncertainty expressed by head movements only versus by both head and eyebrow movements. On the other hand, it contradicts the findings in the case of certainty, where certainty expressed by eyebrows was found to be effective, and even more so in combination with head movements. To sum up, there seem to be no straightforward alternative hypotheses.

With hindsight, the experimental setup has a number of weaknesses that should be properly addressed in future work. One of these is the simplifying assumption that the expected mean score is equal to the mid of the scale (3). However, answers may be inherently more certain or uncertain because of their semantic content. This inherent bias can be measured by running a separate experiment in which subjects are asked to rate certainty on the basis of the text only. This bias can then be taken into account during analysis and statistical testing.

Another issue is that the question *How certain do you think the speaker is of the provided answer?* severely constraints the range of responses. Without this strong bias, subjects might prefer to interpret the facial expressions along other, unintended dimensions such as *surprise* or *agitation*, rather than *certainty.* One possible method to reduce this bias is to ask subjects to score on other scales besides the one for certainty.

We found there is a tension between RUTH's (theoretical) requirement that batons should be time aligned with accented words and that of a natural rendering of facial movements. Our animations frequently had batons at unaccented words. Moreover, the recommendation that underliners should be aligned with the phrasal tones of intermediary phrases is even impossible in Dutch, as there is no such thing in descriptions of Dutch intonation. This suggests more research is needed on the topic of alignment between intonational and facial movements.

In order to keep the experiment manageable, we limited ourselves to eyebrow and head movement. However, RUTH supports at least two other movements: smiles and blinks. It would be interesting to run a similar experiment using these cues. At the same time, the repertoire of current talking heads is much more constrained than that of real humans. For instance, Swerts and Krahmer (2005) mention a complex expression they labeled *funny face*, which their subjects often used to express uncertainty.

4 General Discussion and Conclusion

In order to retain a user's trust, QA systems need to express the level of uncertainty attached to their answers. Multimodal QA systems offer the opportunity to express uncertainty through other than verbal means. On the basis of evidence from studies how uncertainty is expressed in human-human dialogue, it was argued that uncertainty is better expressed by audiovisual than by verbal means. Moreover, we summarized (unpublished) work on visual expression of uncertainty in the context of QA systems suggesting that humans dislike linguistic signaling of uncertainty and prefer visual signaling instead. Circumstantial evidence comes from general work on trust and ECA's.

An experiment was described to test whether we can reliably express certainty or uncertainty by means of a limited repertoire of animated facial expressions, in particular, only combinations of eyebrow movements and head movements. The results suggest that humans can correctly recognize animated facial expressions as certain, but that only head movements are a consistent cue. We discussed a number of issues with the experimental setup which preclude definite conclusions.

In addition, there are a number of open issues that need to be resolved before a talking head like RUTH can be employed for signaling uncertainty in multimodal human-computer interaction. If we take the IMIX multimodal QA system as a case in point, it is assumed that its QA engines can provide reliable confidence scores. In practice, however, it turns out that it is hard for a system to know that is does not know the answer, let alone how certain it is of a particular answer. Future development in QA is likely to improve this (Burger et al., 2003).

It should also be noted that our results only concern two extremes, i.e. certainty versus uncertainty. In a practical system, a more likely setting is to express a *degree* of certainty. Our results in part suggest that a combination of cues gives a stronger effect, but more research is definitely required.

Another open issue is how to obtain the specifications for facial expressions. So far our annotations were produced manually, but a dialogue system should of course be able to predict them automatically. For some limited domains the use of templates may be sufficient, but in QA systems like the IMIX system, where text variation is unpredictable, such an approach is unlikely to succeed. Given the similarity to the problem of predicting prosodic markup in speech synthesis, and the successful application of machine learning techniques in that area (e.g. Marsi et al., 2003), a data-driven approach seems most promising. For training and evaluation purposes, we would then need a substantial corpus of annotated examples – of either human speakers or ECA's – and select informative (linguistic) features. One of our own topics for future research is data-driven prediction of annotations to appropriately express uncertainty.

References

Black, A. W., Taylor, P., and Caley, R. (2002). *The Festival Speech Synthesis System, System documentation*. Centre for Speech Technology Research University of Edinburgh.

Boves, L. and den Os, E. (2005). Interactivity and multimodality in the IMIX demonstrator. In *International Conference on Multimedia and Expo*, pages 1578–1581.

Brennan, S. and Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34(3):383–398.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C., Maiorano, S., Miller, G., et al. (2003). Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). Technical report.

Burgoon, J. (1994). Nonverbal signals. In Knapp, M. L. and Miller, G. R., editors, *Handbook of Interpersonal Communication*, volume 2, pages 229–285. Sage.

Cassell, J. and Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of ACM*, 43(12):50–56.

Chovil, N. (1991a). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194.

Chovil, N. (1991b). Social determinants of facial displays. *Journal of Nonverbal Behaviour*, 15(3):141–154.

DeCarlo, D. and Stone, M. (2003). The Rutgers University Talking Head: RUTH. Technical report, Rutgers University.

DeCarlo, D., Stone, M., Revilla, C., and Venditti, J. (2004). Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38.

Gussenhoven, C. (2005). Transcription of Dutch intonation. In Jun, S.-A., editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, Oxford.

Hart, J. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56:208–216.

Krahmer, E. and Swerts, M. (2005). How children and adults signal and detect uncertainty in audiovisual speech. *Language and Speech*, 48(1):29–54.

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. (2003). What makes a good answer? The role of context in question answering. In *Human-Computer Interaction (INTERACT 2003)*.

Marsi, E., Busser, G., Daelemans, W., Hoste, V., Reynaert, M., and van den Bosch, A. (2003). Learning to predict pitch accents and prosodic boundaries in Dutch. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 489–496, Sapporo, Japan.

McClave, E. (2000). Linguistic functions of head movement in the context of speech. *Journal of Pragmatics*, 32:855–878.

Moldovan, D., Clark, C., and Harabagiu, S. (2003). COGEX: a logic prover for question answering. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 87–93. Association for Computational Linguistics Morristown, NJ, USA.

Poggi, I. (2002). Towards the alphabet and lexicon of gesture, gaze and touch. In Bouissac, P., editor, *Virtual Symposium on Multimodality of Human Communication.* http://www.semioticon.com/virtuals/index.html.

Riegelsberger, J., Sasse, M., and McCarthy, J. (2005). Do people trust their eyes more than their ears?: Media bias in detecting cues of expertise. In *Conference on Human Factors in Computing Systems*, pages 1745–1748. ACM Press New York, NY, USA.

Smith, V. and Clark, H. (1993). On the course of answering questions. *Journal of Memory and Language*, 32:25–38.

Swerts, M. and Krahmer, E. (2005). Audiovisual prosody and Feeling of Knowing. *Journal of Memory and Language*, 53(1):81–94.

Swerts, M., Krahmer, E., Barkhuysen, P., and van de Laar, L. (2003). Audiovisual cues to uncertainty. In *Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pages 14–3, Chateau-D'Oex, Switzerland.

Theune, M., van Schooten, B., op den Akker, R., Bosma, W., Hofs, D., Nijholt, A., Krahmer, E., van Hooijdonk, C., and Marsi, E. (to appear 2007). Questions, pictures, answers: Introducing pictures in question-answering systems. In *Proceedings of the Tenth International Symposium on Social Communication*, Santiago de Cuba, Cuba.

Voorhees, E. (2003). Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC2003).*

Zaihrayeu, I., da Silva, P., and McGuinness, D. (2005). IWTrust: Improving User Trust in Answers from the Web. In *Proceedings of 3rd International Conference on Trust Management (iTrust2005)*, Rocquencourt, France. Springer.

Vocabularies for Description of Accessibility Issues in Multimodal User Interfaces

Željko Obrenović, Raphaël Troncy, Lynda Hardman
CWI, P.O. Box 94079, 1090 GB, Amsterdam, The Netherlands
Firstname.Lastname@cwi.nl

Abstract

In previous work, we proposed a unified approach for describing multimodal human-computer interaction and interaction constraints in terms of sensual, motor, perceptual and cognitive functions of users. In this paper, we extend this work by providing formalised vocabularies that express human functionalities and anatomical structures required by specific modalities. The central theme of our approach is to connect these modality representations with descriptions of user, device and environmental constraints that influence the interaction. These descriptions can then be used in a reasoning framework that will exploit formal connections among interaction modalities and constraints. The focus of this paper is on specifying a comprehensive vocabulary of necessary concepts. Within the context of an interaction framework, we describe a number of examples that use this formalised knowledge.

Keywords: Multimodal interaction, universal accessibility, inclusive design, formal models

1 Introduction

The long-term goal of our research is to use formal models of multimodal user interfaces and interaction constraints to allow the (semi-)automatic analysis of required human functionalities and anatomical structures for a particular (multimodal) interface. Figure 1 illustrates the basic theoretical framework for our approach: we describe multimodal user interfaces as systems that communicate a message, an effect, by means of a modality stimulating a particular human functionality or anatomical structures, such as, sensory, motor, perceptual or cognitive. On the other hand, constraints describe influence on various factors on human anatomical structures and functionalities. For example, a simple text presentation engages many visual perceptual functions, such as shape recognition, visual grouping by proximity, grouping by good continuation, as well as other cognitive and linguistic functions. Interaction constraints, such as user disability or environmental conditions, reduce or completely eliminate some of the effects. For example, users with a central field loss disability cannot read text at usual font sizes in usual lighting conditions. By combining these descriptions, it is possible to see if the designed interface will be appropriate for a specific situation, and it enables adaptation of user interfaces according to user profiles and situational parameters. With our approach, developers can concentrate on more generic effects, providing solutions for different levels of availability of specific functionalities or anatomical structures. In this way, it is possible to create adaptable solutions that adjust to user features, preferences and environmental characteristics, Obrenovic and Starcevic (2004); Obrenovic et al. (2007).

From a developer's point of view, an advantage of this framework is that it is possible to design more flexible and more reusable solutions, aimed at a broader set of situations. Most previous work on designing solutions for people with disabilities focuses on a specific set of disabilities, or on specific situations, Abascal (2002). Bearing in mind the diversity of disabilities and situations, it is clear that development and maintenance of such systems is complex and non-optimal. An advantage of a unified description of user features, preferences and environmental characteristics, is the potential for reusing solutions, created for a particular disability, for non-disabled users in situations that limit the interaction in the same way. As well as providing more universal

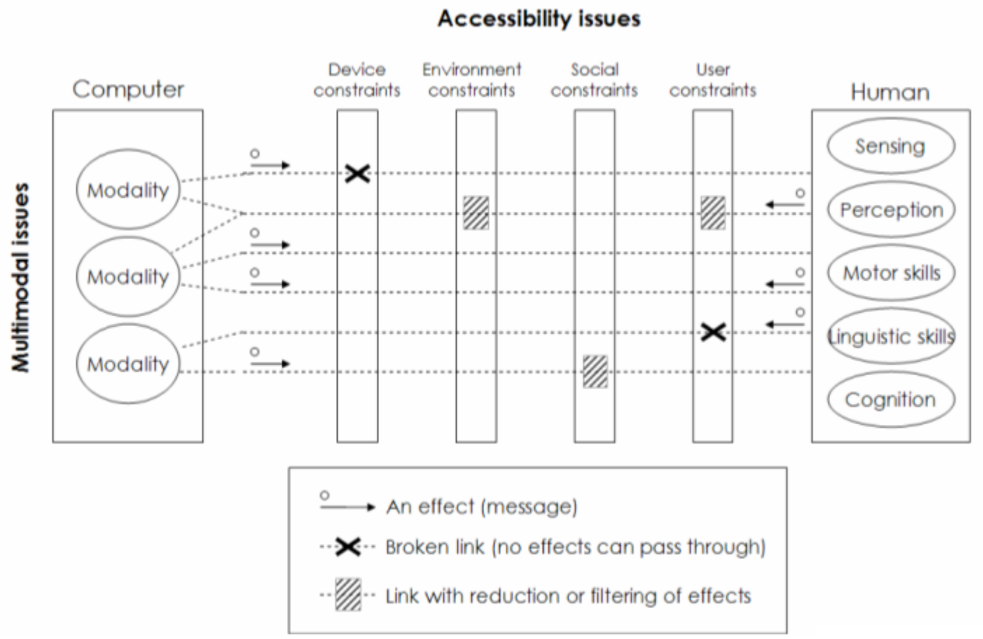


Figure 1: Modalities, constraints, and effects from Obrenovic et al. (2007). Computers and humans establish communication channels over which they exchange messages (effects) that engage a subset of human functionalities and anatomical structures. Modalities produce these effects, while various interaction constraints reduce or completely eliminate some of these effects.

solutions, this could also solve a number of ethical problems, since the design concerns effects and their constraints, rather than the term 'disability', which often introduces negative reactions. Indeed, constraints are often not a consequence of user physical limitations. For example, when interacting with a computer while driving a car, the driver is in a similar situation as a user with limited vision. These situations thus do not have to be treated differently, and solutions from one domain can be reused in another domain.

In the following, we present the basic idea of our approach, and discuss the main topic of the paper: the definition of a comprehensive set of vocabularies as a formal description of modalities and constraints (section 2). We present then some simple use cases where we have used terms from our framework to describe a concrete user interface and the human functionalities and anatomical structures required (section 3). Finally, we conclude the paper and outline some future work (section 4).

2 Vocabularies for Describing Accessibility Issues

A central problem for describing accessibility issues in multimodal interfaces is the definition of a vocabulary for the description of interaction effects in terms of human functionalities. Such a vocabulary would provide terms for describing abstract models of multimodal interaction. In this respect, our approach is similar to existing work in the area of abstract user interface representations, such as User Interface Markup Language (UIML), Extensible Interface Markup Language (XIML), W3C XForms and Alternate Interface Access Protocol (AIAP), Trewin et al. (2003). These abstract models define a vocabulary of modeling primitives for describing elements of user interfaces. Several research groups have tried to improve Web accessibility by adding annotations to web pages to help users understand the meaning of the information as opposed to its presentation and order, Bechhofer et al. (2006). Researchers have also emphasised the importance of user information and its relationships to device profiles, Velasco et al. (2004), discussing vocabularies

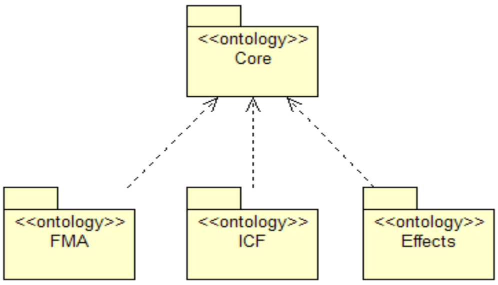


Figure 2: Ontologies for Describing Accessibility Issues

that should be used for description of these profiles. However, many of these solutions mostly focus on abstracting existing user interface platforms and content description, i.e. they are closer to implementation technology, and they do not provide a vocabulary for describing important accessibility issues and human factors involved in interaction. Our goal is to add semantics about accessibility issues and human factors. A similar attempt to defining vocabularies used for description of multimodal interaction has been taken by Ole Bernsen (1994). In his Modality theory, he introduces a generative approach to the analysis of modality types and their combinations, based on his taxonomy of generic unimodal modalities of representation. In this theory, each interaction modality is described in terms of five properties: linguistic (yes/no), analog (yes/no), arbitrary (yes/no), dynamic (yes/no) and media (visual/audio/haptic). In our work, however, we advocate a more generic solution, which enables describing human factors involved in multimodal interaction with more details and with standard vocabularies.

Instead of creating a new vocabulary from scratch we exploit two existing resources (Figure 2): *The International Classification of Functioning, Disability and Health* (ICF)[1] and *The Foundational Model of Anatomy* (FMA)[2]. While these resources cover the description of human functionalities and human anatomy comprehensively, they lack descriptions of interaction effects at the required level of granularity. To compensate for this, we also propose our own vocabulary for describing interaction effects in a multimodal environment.

These separate vocabularies need to be conceptually integrated as well as expressed in a language that can be processed within a system. The main contribution of this paper is the formalisation and conceptual integration of these three resources. We use the Web Ontology Language (OWL)[3] as the main language for describing these vocabularies, which allows us to use the large number of existing suites of knowledge management technologies and tools. In the following, we first describe each of the three vocabularies and then discuss how they can be combined.

2.1 The International Classification of Functioning, Disability and Health (ICF)

The International Classification of Functioning, Disability and Health (ICF) defined by the World Health Organisation provides a comprehensive overview of many important functions of humans. The ICF is a good candidate for describing all important human functionalities, as it provides a detailed description of human functions structured around the following broad components:

- body functions and structure,
- activities (related to tasks and actions by an individual) and participation (involvement in a life situation), and
- information on severity and environmental factors.

[1] `http://www3.who.int/icf/onlinebrowser/icf.cfm`
[2] `http://sig.biostr.washington.edu/projects/fm/`
[3] `http://www.w3.org/2004/OWL/`

In ICF, functioning and disability are viewed as a complex interaction between the health condition of the individual and the contextual factors of the environment as well as personal factors. The picture produced by this combination of factors and dimensions is of "the person in his or her world". The classification treats these dimensions as interactive and dynamic rather than linear or static. ICF has, however, several shortcomings that complicate its formalisation, including[4]: contrasting classifications, confusion between classes of activities and their qualities or features, incorrect and incomplete classifications, and over-simplification or over-emphasis of parts. Nevertheless, this resource is widely used in the health community, and by providing some connection with it, we are able to use the same standard terminology, and possibly reuse medical profiles described in these terms.

We have formalised part of the ICF ontology as an OWL ontology. Currently we have included only the concepts required by the ICF checklist[5] since they provide a good summary of the content of the whole classification. The formalization of 160 concepts reproduces the *is-a* hierarchy given in the checklist, distributed on four levels where the top-level concepts are `ActivitiesAndParticipations`, `BodyFunctions`, `BodyStructures` and `Environment`.

2.2 The Foundational Model of Anatomy (FMA)

Another useful resource is the Foundational Model of Anatomy (FMA). FMA represents a coherent body of explicit declarative knowledge about human anatomy. It is developed and maintained by the Structural Informatics Group at the University of Washington. It has also been formalised in OWL by the medical informatics group at Stanford. However, although we can directly use FMA, it misses many important concepts about human functionalities, as they cannot be described by anatomical properties. OWL release of FMA is available at: `http://webrum.uni-mannheim.de/math/lski/release.html`.

2.3 Interaction Effects

ICF and FMA provide a number of concepts about human functionalities and anatomy, but they still lack terms for more detailed description of effects that some modalities produce. For example, with ICF, we can specify that an interaction modality requires human visual perception, and FMA can provide us with a description of all parts of the human perceptual system, but none of these resources provides terms for describing details, such as, if it is expected that users perceive grouping, highlighting, or three-dimensional position of the objects. Furthermore, there are also different ways how perceptual grouping, highlighting or three-dimensional perception can be achieved. To overcome this problem, we have created a simple taxonomy of interaction effects not covered by ICF or FMA, Obrenovic and Starcevic (2004). This vocabulary describes additional sensory, motor, perceptual, and cognitive effects, from resources such as Gestalt psychology. We have formalised this resource as an OWL ontology that contains 114 concepts, some of which are shown in Table 1.

2.4 Combining the Vocabularies

In order to use the various vocabularies together, we need to connect them. Wache et al. (2001) reports three ways for doing so, namely the *single ontology*, the *multiple ontologies* and the *hybrid* approaches. In the first approach, all the vocabularies are merged in a single global ontology, while in the second one, an additional representation formalism defining the inter-ontology mapping is needed. The hybrid approach, which we have adopted, considers both aspects: separate vocabularies co-exist and are linked using a core-level ontology.

Figure 3 presents the basic concepts defined in our core ontology, with the relations to the key concepts from the three vocabularies described above. This ontology extends our previous proposal of interaction modalities, Obrenovic and Starcevic (2004), and interaction constraints,

[4] See: `http://ontology.buffalo.edu/medo/ICF.pdf`

[5] `http://www3.who.int/icf/checklist/icf-checklist.pdf`

- Grouping
  - Gestalt visual grouping
    - Grouping by similarity
    - Grouping by motion
    - Grouping by texture
    - Grouping by symmetry
    - Grouping by proximity
    - Grouping by parallelism
    - Grouping by closure
    - Grouping by good continuation
- Highlighting
  - Gestalt visual highlighting
    - Highlighting by color
    - Highlighting by polarity
    - Highlighting by brightness
    - Highlighting by orientation
    - Highlighting by size
    - Highlighting by motion
    - Highlighting by flicker
    - Highlighting by depth
    - Highlighting by shape
  - Audio highlighting
    - Highlighting by intensity
    - Highlighting by pitch
    - Highlighting by rate
- 3D cue
  - Visual 3D cues
    - Stereo vision
    - Motion parallax
    - Linear perspective (converting lines)
    - Relative size
    - Shadow
    - Familiar size
    - Interposition
    - Relative height
    - Horizon
  - Audio 3D cues
    - Inter-aural time (or phase) difference
    - Inter-aural intensity (or level) difference
    - Head Related Transfer Functions (HRTFs)
    - Head movement
    - Echo
    - Attenuation of high frequencies

Table 1: Some perceptual effects defined in the interaction effects ontology

Obrenovic et al. (2007). The integration of the three vocabularies described above with our core ontology is available at: `http://www.cwi.nl/~media/ontologies/multimodality.owl`.

We introduce the concept of human entity, which describes an anatomical structure, or a function. An interaction modality can then be described in terms of the human entity it requires for interaction. An interaction constraint is defined in terms of the human entities that it restricts. The FMA ontology provides a number of concepts for describing human anatomical structures. The ICF body structure concepts provide a similar, but less detailed, classification of human anatomical structures. The FMA and ICF body structure concepts overlap, but FMA provides much more comprehensive and better formalised data. For the description of human functional artifacts, ICF provides concepts for the description of body functions, and functions related to human activity and participation. Neither of these, however, allows more detailed description of many parts. Our interaction effects ontology fills this gap by defining additional functional entities at sensory, perceptual and cognitive levels. These three vocabularies together provide sufficient coverage of human functionalities and effects to allow the types of mappings we envisage between the available functionalities and appropriate modalities.

In addition to this coverage of description, we also need mappings between the different vocabularies. Currently, the only relation among concepts from different vocabularies is the human entity concept. For a more elaborate analysis, more relations among concepts are necessary, for example, by establishing a mapping between the ICF body structure and the FMA concepts. These relations are also necessary to enhance the description. For example, if we describe that a user is not able to process a sound, it means that not only the sensory, but also all the audio perceptual effects will not be appropriate for that user. If the user cannot use the central visual field, limitation of vision processing will affect all visual perceptual effects, as well as linguistic effect of reading. In a similar way, low colour processing will decrease the use of the highlighted colour effect, while contrast processing will reduce shape recognition and highlighting by brightness.

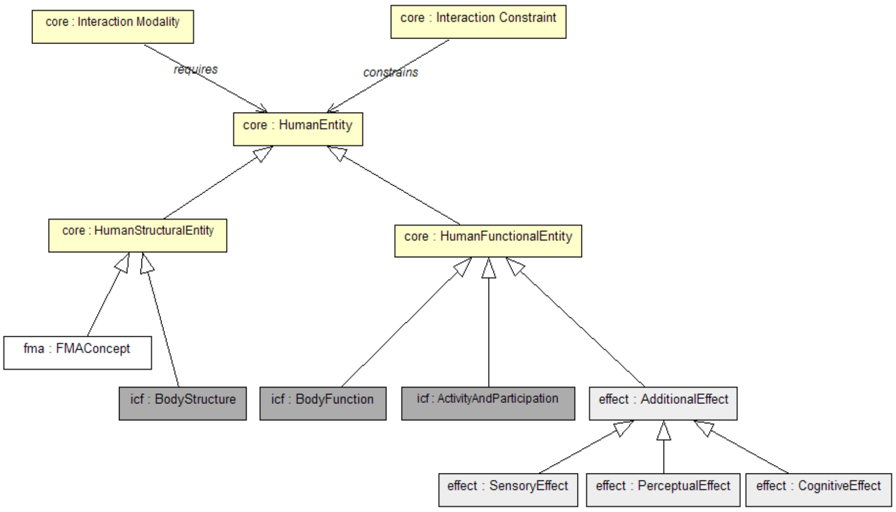

core : Interaction Modality
core : Interaction Constraint
requires
constrains
core : HumanEntity
core : HumanStructuralEntity
core : HumanFunctionalEntity
fma : FMAConcept
icf : BodyStructure
icf : BodyFunction
icf : ActivityAndParticipation
effect : AdditionalEffect
effect : SensoryEffect
effect : PerceptualEffect
effect : CognitiveEffect

Figure 3: The core ontology, with relations to the key concepts from the three vocabularies

3 Describing Accessibility Issues with the Vocabularies

In this section we present a number of examples that illustrate how we can use the vocabularies to describe accessibility issues in multimodal user interfaces. First we show how the vocabularies can be used to describe the requirements of standard interaction modalities. We then show an example description of (implicit) design decisions. Finally, we illustrate descriptions of interaction constraints.

3.1 Describing Interaction Modalities

We describe interaction modalities in terms of the human entities they require in order to enable interaction. These descriptions can provide richer semantics about many implicit requirements of interaction modalities. For example, figure 4 shows a simplified description of human entities required by the aimed-hand movement, a modality often used in graphical user interfaces. Aimed-hand movement is a complex modality that integrates hand movement input with visual feedback. To describe these modalities, we need concepts from all three vocabularies. A hand movement input modality, such as that used to control the mouse, requires human hand anatomy (described with concepts from the FMA ontology), no impairments in the mobility of joints, muscle power and muscle tone, plus an absence of involuntary movements (described with concepts from the ICF ontology). Visual feedback requires user eye and visual context (concepts from FMA ontology) and seeing and attention functionalities (concepts from ICF ontology). In addition we describe additional perceptual functions introduced by visual feedback: highlighting by motion and shape of the cursor, and optionally with depth if the cursor has shadow (these concepts are defined in the interaction effects ontology).

Figure 5 shows a simplified description of interaction requirements of speech interaction. This is a complex modality that integrates speech input and output (defined relative to the computer). On an anatomical level (described with FMA concepts) speech interaction requires human vocal tract (stomatognathic system) for user speech, ear, auditory cortex and auditory additional cortex. On a functional level (described with ICF concepts) speech interaction requires functions of speaking, voice, hearing, receiving spoken message, language, conversational, general voice and speech functions, and usage of short term memory.

More complex modalities, such as those that use three-dimensional presentation, can also be described in this way, Obrenovic and Starcevic (2004). We can combine these descriptions with descriptions of interaction constraints, such as those described in section 3.3, to see if particular

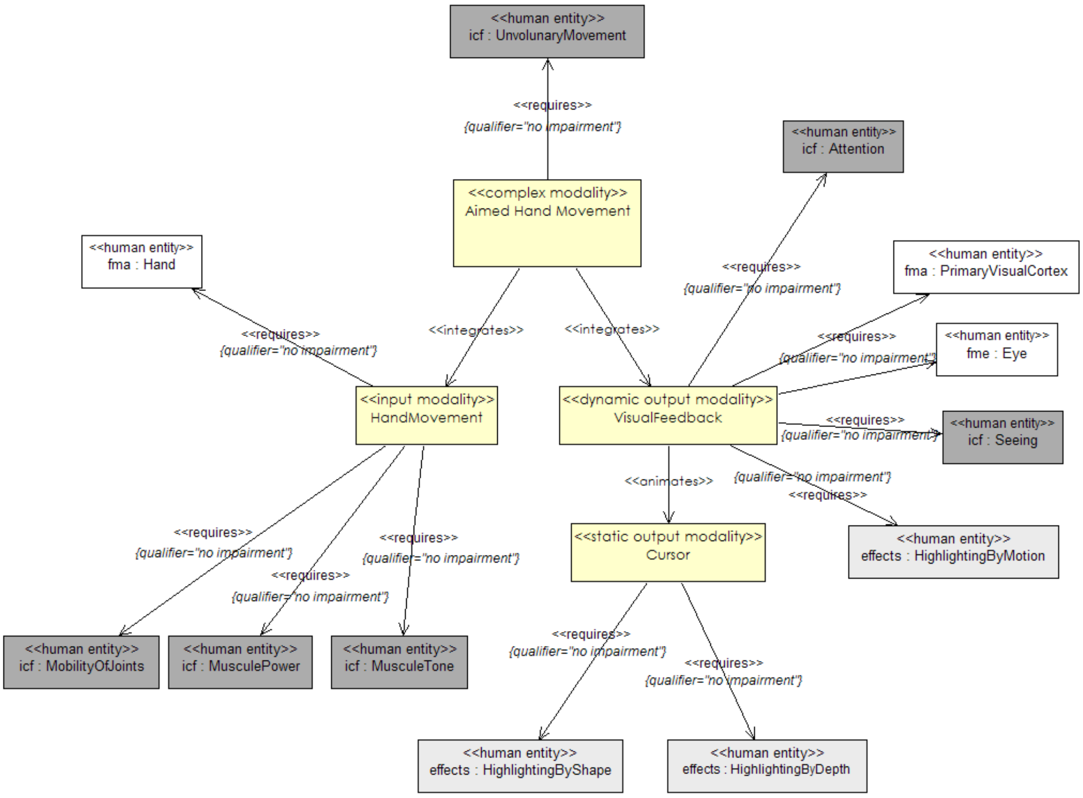


Figure 4: Description of human entity required by aimed hand movement

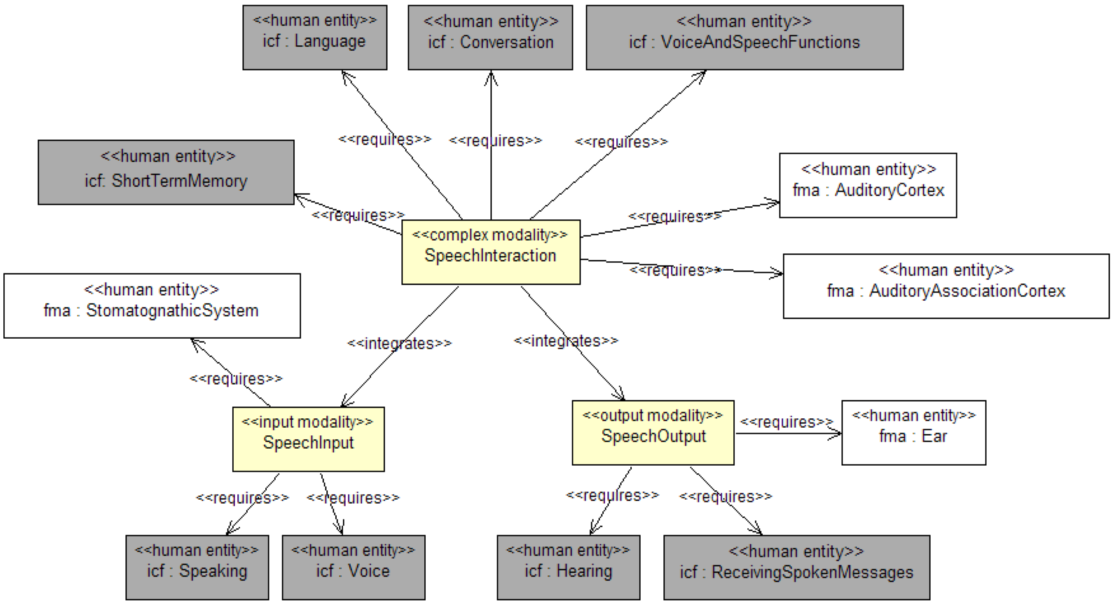
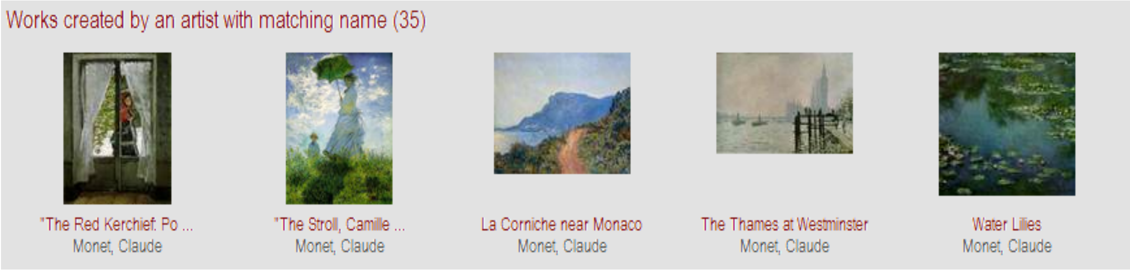


Figure 5: Description of human entities required speech-interaction



Works created by an artist with matching name (35)

"The Red Kerchief: Po ...
Monet, Claude

"The Stroll, Camille ...
Monet, Claude

La Corniche near Monaco
Monet, Claude

The Thames at Westminster
Monet, Claude

Water Lilies
Monet, Claude

Figure 6: Simplified images of a part of MultimediaN e-culture interface

modalities can be used in a given context. We can also use descriptions of interaction modalities to identify potential conflicts in requirements. For example, Karl et al. (1993) found that the use of speech to issue commands interfered with short-term memory requirements that constituted part of the experimental task. Applying our proposed modeling of combinations of modalities allows us to select those with non-conflicting requirements.

3.2 Describing Designers' (Implicit) Decisions

User interfaces can be viewed as one-shot, higher-order messages sent from designers to users, Prates et al. (2000). In designing a user interface, the designer defines an interactive language that determines which messages will be included in the interaction. However, multimodal user interfaces are usually implemented with commercially available implementation platforms that do not integrate the concepts of modality and multimodal integration. As a result, it can be impossible to determine the designer's original intent, which can be important when analysing and reusing parts of the user interface. The vocabularies that we have presented can be used to describe some of these intentions, enabling a designer or an HCI expert to state their aims and accessibility requirements of the interface.

Figure 7 shows an example of how we can describe the design intentions of a particular interface. The figure focuses on the descriptions of perceptual effects used in a user interface of the MultimediaN E-Culture project[6] shown in Figure 6. The interface shows an ordered list of images, with their associated titles and names of artists. Even though we describe a simple part of the interface, there are many important implicit elements of this presentation. Images, image titles, and artist names are perceptual entities grouped by proximity, in order to be perceived as a whole. The image titles and artist names are linguistic modalities, requiring user reading and knowledge of language in order to be understood. The image title is also a hyperlink, visually highlighted by a colour, and by flicker when the mouse cursor is moved over it. Image presentations are grouped in a horizontal line in order to exploit perceptual effect of grouping by good continuation, and by similarity of their shape. The images are sorted so that the user exploits left-to-right perception of ordering.

This example shows many high-level effects used in the interface. With ICF and FMA we can only say things that are common for graphical user interfaces, i.e that the interface requires a human eye, a visual cortex, user visual perception and the function of reading. This example thus illustrates the need for concepts not present in ICF and FMA.

3.3 Describing Interaction Constraints

Constraints are associated with a set of human entities that they restrict. As we describe interaction modalities and constraints using the same vocabularies, we can combine these descriptions with descriptions to identify potential interaction problems or select modalities that are not affected by the constraints.

[6] `http://e-culture.multimedian.nl/demo/search`

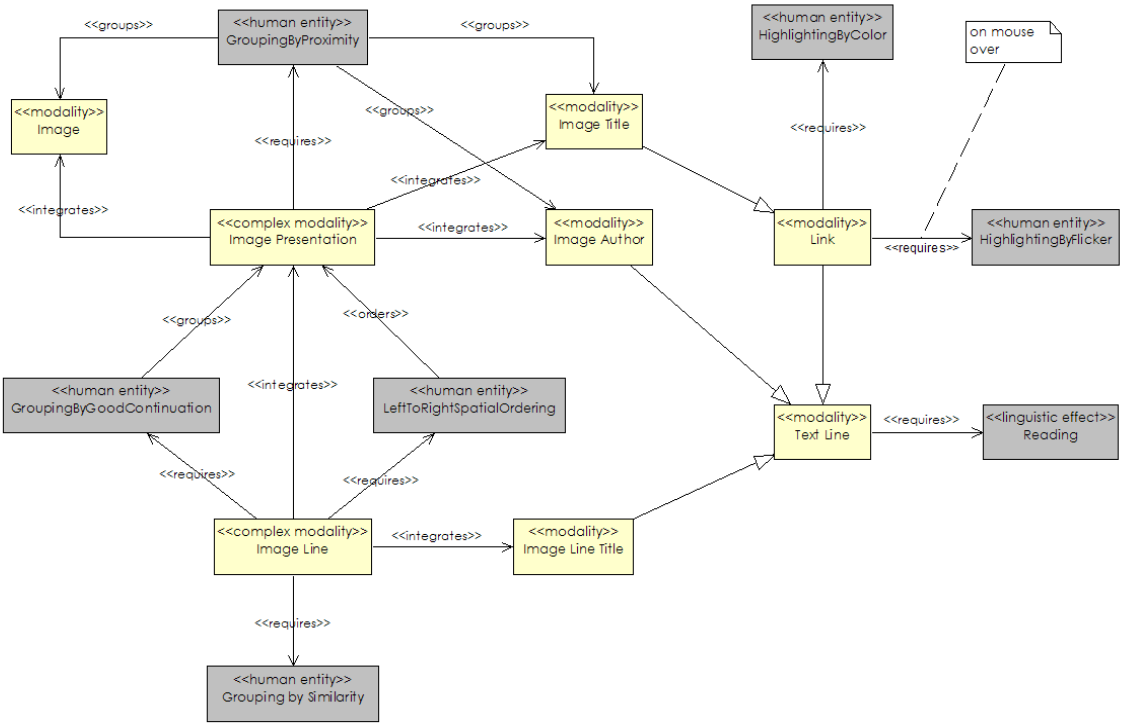

<<human entity>> GroupingByProximity
<<groups>>
<<groups>>
<<human entity>> HighlightingByColor
on mouse over
<<modality>> Image
<<groups>>
<<modality>> Image Title
<<requires>>
<<requires>>
<<integrates>>
<<integrates>>
<<complex modality>> Image Presentation
<<integrates>>
<<modality>> Image Author
<<modality>> Link
<<requires>>
<<human entity>> HighlightingByFlicker
<<groups>>
<<orders>>
<<integrates>>
<<human entity>> GroupingByGoodContinuation
<<human entity>> LeftToRightSpatialOrdering
<<modality>> Text Line
<<requires>>
<<linguistic effect>> Reading
<<requires>>
<<requires>>
<<complex modality>> Image Line
<<integrates>>
<<modality>> Image Line Title
<<requires>>
<<human entity>> Grouping by Similarity

Figure 7: UML diagram describing modalities and effects used in the E-Culture interface

Figures 8 and 9 show UML models of two interaction constraints: central field loss disability, and noise.

Central field loss is a disability that limits fovea processing to a very low level (Figure 8). In terms of FMA, this means that this modality constrains the fovea centralis element. In terms of human functions, the disability constrains ICF functions of seeing, watching, reading, writing, learning to read and write, and, indeed, many perceptual functions in general.

Noise in the environment primarily constrains human speech and audio interaction (Figure 9). In terms of FMA, this modality constrains effects associated with the ear. In terms of human functions, it affects ICF functions of speaking, hearing, receiving spoken messages, conversation, attention, and interaction effects of audio highlighting and grouping.

4 Conclusion and Future Work

There are many steps that have to be taken to achieve our long-term goal of using formal models of interaction modalities and interaction constraints to build solutions that can automatically analyse accessibility issues. A first step is the definition of a comprehensive vocabulary for formally describing modalities and constraints. When such a vocabulary exists, even in a simple form, it is possible to improve the design of multimodal user interfaces in many directions. The main benefit of models created with such a vocabulary is an explicit representation of accessibility issues using widely understood terms. Explicit representation leads to more automation, while using standards for knowledge representation, we automatically inherit the potential for reuse of many existing knowledge analysis tools.

Our next step is the definition of a reasoning framework that can exploit the semantics from descriptions of interaction modalities and constraints (Figure 10). The basic idea of the framework is that applications define the context by providing descriptions of user interfaces in terms of interaction constraints, and description of user, device and environment profiles in terms of interaction

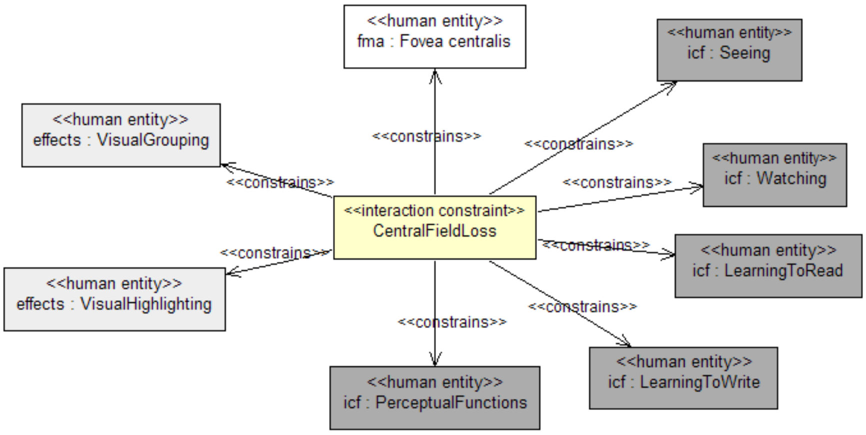


Figure 8: Description of central field loss disability (fovea vision loss)

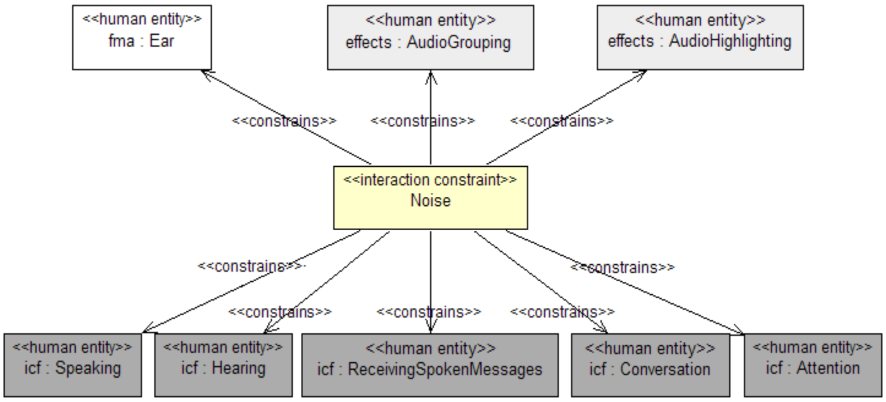


Figure 9: Description of influence of noise

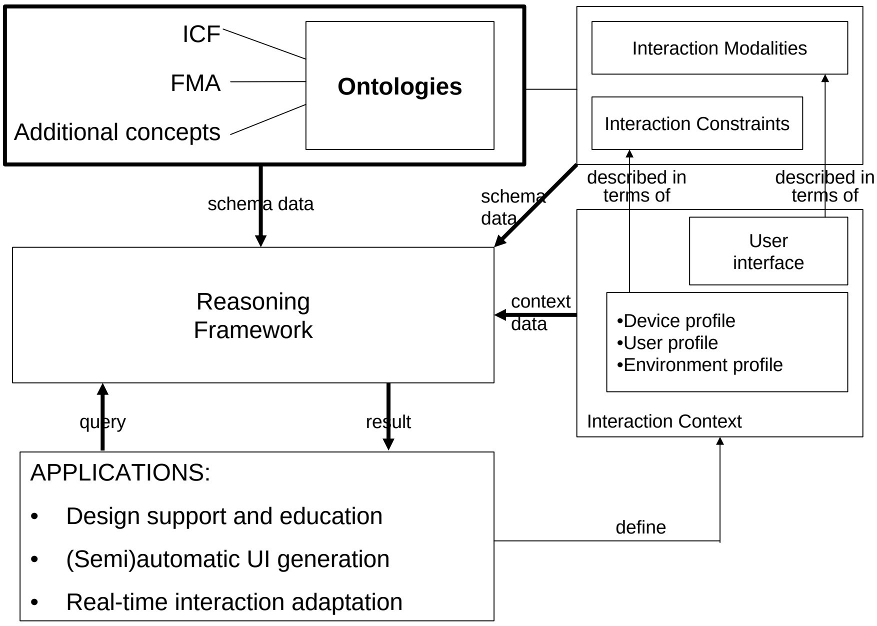


Figure 10: Using the descriptions of accessibility issues

constraints, and then use the framework to reason over these data and semantics relations. The reasoning framework can be used as a design support and education tool, enabling designers to verify their high-level decisions, and explore relations among concepts. Systems that generate user interfaces can use it to select appropriate modalities, or change them in real-time. The proposed framework can also be a good basis for approaches, such as user interface adaptation and content repurposing, that tackle the problem of developing content for various users and devices. The main idea of our approach is that existing content can be analysed in order to create higher-level descriptions using the concepts from our ontologies. If original content is not appropriate for the user or situation, we can try to repurpose it into a new form, changing inappropriate modalities, while maintaining the higher-level effects contained in the user interface.

Acknowledgments

The research leading to this paper was partially supported by the European ITEA Passepartout Project and by the European Commission under contract FP6-027026, "Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space". We would like to thank Alia Amin, Joost Geurts and Jacco van Ossenbruggen who provided useful feedback on the work described here and whose suggestions significantly improved this article.

References

Abascal, J. (2002). Human-computer interaction in assistive technology: from "Patchwork" to "Universal Design". In *IEEE International Conference on Systems, Man and Cybernetics.*

Bechhofer, S., Harper, S., and Lunn, D. (2006). SADIe: Semantic Annotation for Accessibility. In *5th International Semantic Web Conference (ISWC'06)*, pages 101–115, Athens, GA, USA.

Bernsen, N. O. (1994). Foundations of Multimodal Representations: A Taxonomy of Representational Modalities. *Interacting with Computers*, 6(4):347–371.

Karl, L., Pettey, M., and Shneiderman, B. (1993). SpeechActivated versus Mouse-Activated Commands for Word Processing Applications: An Empirical Evaluation. *International Journal of Man-Machine Studies*, 39(4):667–687.

Obrenovic, Z., Abascal, J., and Starcevic, D. (2007). Universal accessibilty as a multimodal design issue. *Communications of ACM (to appear)*.

Obrenovic, Z. and Starcevic, D. (2004). Modeling Multimodal Human-Computer Interaction. *Computer*, 37(9):65–72.

Prates, R. O., de Souza, C. S., and Barbosa, S. D. J. (2000). Methods and tools: a method for evaluating the communicability of user interfaces. *Interactions*, 7(1):31–38.

Trewin, S., Zimmermann, G., and Vanderheiden, G. (2003). Abstract User Interface Representations: How Well do they Support Universal Access? In *2nd ACM Conference on Universal Usability (CUU'03)*, pages 77–84, Vancouver, BC, Canada.

Velasco, A., Mohamad, Y., Gilman, S., Viorres, N., Vlachogiannis, E., Arnellos, A., and Darzentas, S. (2004). Universal access to information services - the need for user information and its relationship to device profiles. *Universal Access in the Information Society*, 3(1):88–95.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-Based Integration of Information - A Survey of Existing Approaches. In *Workshop on Ontologies and Information Sharing*, pages 108–117.

Modality Choice for Generation of Referring Acts[*]
Pointing versus Describing

Paul L.A. Piwek
NLG group, Centre for Research in Computing
The Open University, Walton Hall, Milton Keynes, UK
p.piwek@open.ac.uk

Abstract

The main aim of this paper is to challenge two commonly held assumptions regarding modality selection in the generation of referring acts: the assumption that non-verbal means of referring are secondary to verbal ones, and the assumption that there is a single strategy that speakers follow for generating referring acts. Our evidence is drawn from a corpus of task-oriented dialogues that was obtained through an observational study. We propose two alternative strategies for modality selection based on correlation data from the observational study. Speakers that follow the first strategy simply abstain from pointing. Speakers that follow the other strategy make the decision whether to point dependent on whether the intended referent is in focus and/or important. This decision precedes the selection of verbal means (i.e., words) for referring.

Keywords: generation of referring expressions, choice and integration of output modalities, cognitive aspects, pointing

1 Introduction

In the field of Natural Language Generation, *referring expressions* are defined as '[...] phrases that identify particular domain entities to the human recipient of the generation system's output' (Dale and Reiter, 1995). More precisely, referring expressions are used in *referring acts* to identify domain entities. *Multimodal referring acts* combine verbal means for referring (the aforementioned phrases) with non-verbal means such as pointing. The main aim of this paper is to challenge two assumptions that are commonly held in the literature (see Section 2.1) on generation of multimodal referring expressions:

> (A1) Non-verbal means of referring are secondary to verbal means and only to be resorted to when verbal means are judged inadequate.
>
> (A2) There is a *single* strategy for the choice of output modality when generating of referring acts.

We proceed as follows. Firstly, in Section 2, we discuss and compare several existing approaches to the generation of multimodal referring acts. We also consider some arguments for and against developing algorithms that model human production of multimodal referring acts. In the next section, we examine the existing approaches in the light of data from an observational study, focusing on assumptions A1 and A2. Our own method for modality choice is presented in Section 4. Based on the observational study, we will distinguish between two strategies for choice of output modality. We will also argue that both *domain focus* and *importance* of the intended referent are principal factors that influence modality choice. We conclude this paper with Section 5, where we present our conclusions and some issues for further research.

[*] I would like to thank the three anonymous reviewers of the MOG workshop for their extremely helpful comments and suggestions.

2 Background

Before we proceed, let us lay out the basic assumptions underlying the current study. We focus on referring acts that are produced by a speaker and intended to be understood by an addressee. A referring act is understood if the addressee identifies the object that the speaker had in mind. We focus on situations where several objects are present in the domain and visible to both speaker and addressee. Only one of these objects is the *intended referent* (i.e., we concentrate on singular reference); the other objects are known as *distractors.* The speaker can use spoken language and/or pointing to identify objects in the domain. For instance, when referring to a particular object she might utter 'it', 'that tower with the rounded corners' or 'this tower' accompanied by a pointing act. The main concern of this paper is modality choice, in particular, the decision whether to include a pointing act.

2.1 Related Work

Most work on the generation of multimodal referring acts is based on the assumption that pointing acts are only used if 'proper' verbal means of referring do not suffice. For example, Lester et al. (1999) only include a pointing act, *if* a pronoun cannot be used to refer to the object, and Claassen (1992) goes even further and only resorts to pointing acts when no purely verbal means of identification can be found. Van der Sluis and Krahmer (2001) also see pointing as a last course of action; they use a pointing act only if the object is sufficiently close and a purely verbal referring act would be too complex.

A common thread underlying all these approaches is that the proposed algorithm first tries to formulate an exclusively verbal referring expression, and only if that fails, or the resulting expression becomes to complex, does it start anew and produce a multimodal referring act that includes pointing. This seems like a rather inefficient way to create referring acts, especially given that speakers do often point (see Section 3), and precise pointing is rather effective (since precise pointing rules out *all* distractors at once).

The aforementioned algorithms all focus on precise pointing: the object that is pointed at is uniquely identified by the pointing act. Typically, this is achieved by the (computer-animated) pointing hand touching the object or being very close to it. In contrast, an imprecise pointing act does not single out a unique object, but rather identifies a set of objects (*cf.* Wilkins 1999 and Clark and Bangerter 2004 on close and distant pointing). More recently, some computational work has also considered imprecise pointing. Kranstedt and Wachsmut (2005) speak of object-pointing versus region-pointing. They propose a way – based on a pointing cone originating in the demonstrating hand – for computing whether a particular pointing gesture constitutes precise or imprecise pointing. This then influences the set of distractors that the generation algorithm uses for determining the linguistic content of the referring act. The decision whether or not to point is based on whether the intended referent is visible to both interlocutors: if it is visible, then a pointing act is included. We will, however, see (Section 3) that speakers do not always point even when the intented referent is visible to all interlocutors.

Krahmer and Van der Sluis (2003) propose an account that not only covers different levels of precision of pointing, but also aims to make the decision on whether to point without a priori favouring verbal or non-verbal means of referring. This is achieved by assigning costs to both verbal and non-verbal components of referring expressions. The cost function is set up such that the cost of a non-verbal pointing act is somewhat higher than that of a single verbal component (inclusion of a property such as type or colour). Overall cost is calculated by summing the component costs. Additionally, they introduce three different degrees of precision for pointing: precise, imprecise and very imprecise pointing. The idea is that pointing is a bit like highlighting objects with flashlight: when one is close to the objects it is easy to single out one specific object, whereas as one moves further away, a bigger area is illuminated by the flashlight, thus making it more likely that several objects are highlighted rather than just one. They propose that as precision decreases cost of pointing also decreases.

This approach hinges very much on the values that are assigned by the cost function. Currently, the cost of pointing acts is derived from Fitts's (1954) Index of Difficulty, according to which the difficulty to reach a target is a function of the size of and the distance to a target. This empirically validated index is adapted to the current application by replacing *distance to a target* with *distance from the current position of the hand to the position at which pointing takes place*. Whether this substitution preserves the correctness of Fitts's Index is an open question. More importantly, in addition to the cost for pointing acts, comparable costs for components of referring expressions need to be assigned. Krahmer and Van der Sluis propose without further justification that '[...] type edges (block) are for free, color edges cost 0.75, size edges cost 1.50 and relational edges 2.25.'. In short, a weak aspect of the proposal is that it requires quite a few parameter settings which might be difficult to obtain empirically.[1]

2.2 Automated Generation versus Human Production

All the work we have discussed so far deals with the automated generation of referring acts. In the next section, we examine these algorithms in the light of data from an observational study on multimodal reference by humans. The focus will be on the work by Krahmer and Van der Sluis (2003) which, in our view, represents the most advanced proposal so far. This proposal shares with those by Kranstedt and Wachsmut (2005) and Van der Sluis and Krahmer (2001) the desire to present an algorithm that models human production. The other proposals that we have mentioned are different in this respect.

Lester et al. (1999) introduce the notion of *deictic believability*. Deictic believability is ascribed to a lifelike agent if it simultaneously achieves the following three goals: 1) its spatial references are non-ambiguous, 2) it refers to objects whilst being immersed in the environment (just like human beings can, for example, refer by pointing, speaking and walking at the same time, thus combining gesture, speech and locomotion), 3) its references are pedagogically sound. Both 1) and 3) derive from the learning context for which Lester et al. developed their COSMO system. They are independent of the aim to model human production. Arguably, 2) does suggest that modeling of human production is desirable when developing an algorithm for multimodal generation. Given that human behaviour is generally considered believable (i.e., it generally creates the impression of a sentient being with its own personality and mental states, see Bates 1994), emulating such behaviour could be a good strategy for achieving believability. For Lester et al., believability is, however, not a goal in itself. Rather they argue for it on the basis of the benefits it brings. In particular, they refer to a study (Lester et al., 1997) which showed that believable pedagogical agents are able to produce the *persona effect*: the lifelike character in a learning environment might not have a direct measurable effect on the learning of the students, but it can improve their perception of the learning experience.

The work by Claassen (1992) was done in the context of the EDWARD system. This system was conceived as a prototype multimodal user interface for studying the use and usefulness of multimodal inferfaces. The aim was 'making interaction between a user and a computer more like normal day to day interaction' (Huls and Bos, 1998: 315). Although emulation of human behaviour does not necessarily lead to the most useful systems, it certainly is a good starting point when one is trying to make computer-human interaction more like everyday human-human interaction.

In summary, we have seen that the requirements for certain systems that generate multimodal referring acts are, at least at first sight, independent of the issue of whether or not to model human production. These requirements include *reduction of ambiguity*, *pedagogical soundness* and *interface usability*. It should also be noted that the possibilities afforded by multimodal interfaces can lead to adoption of ways to realize referring acts that do not correspond with human realization. An early example is the CUBRICON system (Neal et al., 1989). This system has the capability to 'point' to the same object simultaneously in different ways: if the object appears in several windows on the computer screen, the strategy is to produce a strong pointing gesture (blinking icon and pointing text box) to the object in the activated window and weak pointing (only highlighting) in all other windows in which the object appears.

[1] Cost functions might, however, have wider applications in the field of referring expressions generation. See, for example, Khan et al. (2006) for further phenomena that may yield to analyses in terms of cost functions.

We have also seen that requirements such as *believability* and *naturalness* do suggest that models of human production of multimodal refential acts are a good starting point for building algorithms, given that human behaviour is our main yardstick for what is considered natural and believable.[2] In this connection, it is also worthwhile to consider an argument that has been put forward by Dale and Reiter (1995:253). This argument suggests that algorithms for generating referring expressions based on psycholinguistic data might be superior to those based on abstract principles (such as the Gricean maxims of conversation). Let us assume that Grice's notion of implicature (Grice, 1975) is correct and can be paraphrased as saying that if a speaker produces an utterance that is unexpected, then the addressee is likely to attempt to infer a reason for why the speaker did not use the expected utterances. Our second premiss is that a system that emulates human behaviour would be more likely to produce expected utterance. If we take these two premisses for granted, then the conclusion follows that systems that are based on a model of real human behaviour are less likely to cause the addressee to erroneously infer unintended reasons for the choice of expression. In contrast, a system based on abstract principles (e.g., avoid ambiguity) might cause the addressee to make inferences purely as result of the unexpected choice of words in the situation at hand (unless, of course, 'avoid ambiguity' is a principle that human speakers tacitly follow anyway).

Last, but not least, let us not forget that even someone who is not persuaded by any of these arguments will hopefully nevertheless concede that computational modeling of human production of referring acts is a valid topic of scientific study in its own right.

3 Empirical Findings from an Observational Study

In this section, we present a number of findings that were derived from a corpus of video-recorded task-oriented spoken dialogues (Cremers, 1996; Beun and Cremers, 1998) obtained in the setting illustrated by Figure 1.[3] The corpus consists of a total of 20 dialogues between Dutch speaking interlocutors. The dialogues arose during a game that the interlocutors were asked to play. In this game, there are two roles, that of the Builder (B), on the right in Figure 1, and that of the Instructor (I). In front of B and I, there is a workspace. The aim of the game is for B to build a structure in the workspace that is a copy of the example structure next to I (on the left in Figure 1). Only I can see the example structure. I and B are, however, allowed to talk with each other and they can also point at the (LEGO) blocks in the workspace. Only B is allowed to move blocks.

For the current study, we used 10 out of the 20 dialogues. At the start of these dialogues, several blocks already occupied the visually shared foundation plate. We examined the initial references to these objects. A total of 121 singular initial references was found after discounting 14 initial plural referring acts, 2 cases of misunderstanding, and 2 cases of self-correction.

Finding 1: *Out of a total 121 singular referring acts in the corpus, 53 included a pointing gesture (Figure 2).* In other words, almost half of the referring acts involved pointing. Such frequent use of pointing suggests that it is more than simply a fall-back strategy for situations where purely verbal referring acts are not adequate.

Finding 2: *The average number of linguistically realized properties in purely verbal referring acts was 1.7, whereas in referring acts that included pointing the average number of properties was 0.98. (two-tailed highly significant at $P \leq 0.0001$, $t = 4.9790$, $df = 119$). See Figure 3 for an overview of the distribution of properties.* This finding is compatible with the Krahmer and Van der Sluis (2003) algorithm; it seems like there is indeed a trade-off between pointing and verbal referring. Additionally, Krahmer and Van der Sluis argue that imprecise pointing is less costly, but also less discriminative. This suggest that it will co-occur with more descriptive content. This could not be verified in the current study, but other work in which distance to target was

[2] The recently emerging user interface paradigm of embodied conversational agents/lifelike characters (Cassell et al. 2004; Prendinger and Ishizuka, 2004) is partly based on the idea that human-computer interaction can be improved by making it more like human-human interaction.

[3] See Cremers (1993) for a written transcript of the corpus.

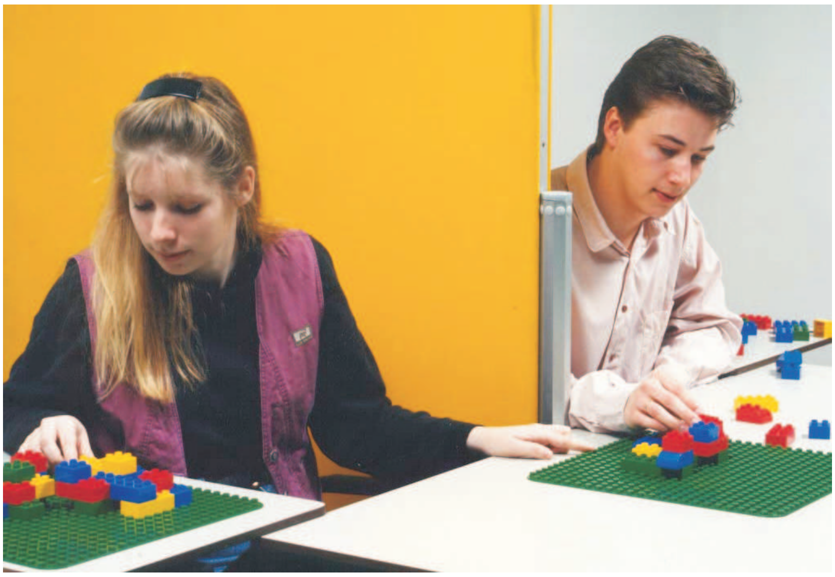


Figure 1: Set-up for collection of task-oriented spoken dialogue corpus involving two roles: Instructor (I) on the left and Builder (B) on the right

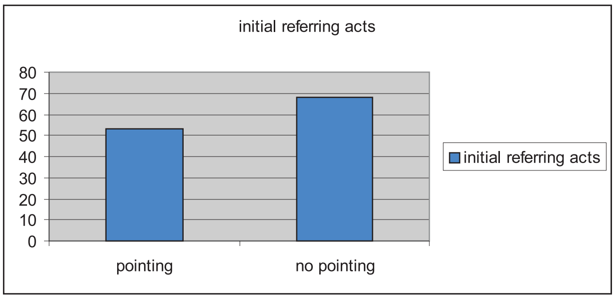


Figure 2: Number of initial referring acts that included and did not include a pointing act.

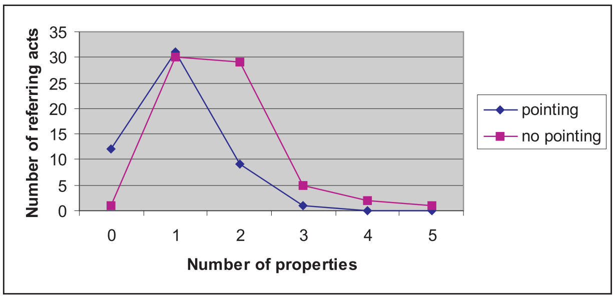


Figure 3: Number of initial referring acts with and without a pointing act mapped against the number of properties expressed by the referring act.

experimentally controlled does seem to confirm this idea (e.g., Bangerter, 2004; Krandstedt et al., 2006).

Finding 3: *Speakers tend to point more frequently when the referent is not in domain focus. When the referent is not in domain focus they point 66% (23/35) of the time, whereas when the referent is in focus they point only 35% (30/86) of the time (significant at $P \leq 0.01$, $\chi^2 = 9.60$, df = 1). See Figure 4.* We discern two types of focus of attention: *discourse focus* (Grosz, 1977) and *domain focus* and concentrate on the latter notion. This notion applies because we are dealing with *initial* identification of objects in a visually accessible domain of discourse. An object is part of the domain focus if it satisfies one of the following two criteria (*cf.* Cremers, 1994):[4]

1. The object was referred to in the preceding utterance or is adjacent to an object that was referred to in the preceding utterance (note that for the initial referring acts that we are considering, it is always the second of these two conditions that is met when an object qualifies as being part of the domain focus);
2. The object lies in an area to which the speaker explicitly directed the attention of the addressee. This is marked by what we will call *focussing expressions* as in 'Wat nou helemaal naar voren zit, daar zit die rode dwars' (literally: *What now entirely to the front is, there is that red one diagonally.* Loose translation: If you look at the bit in the front, you will find a red diagonally placed block).

Finding 3 is also compatible with Krahmer and Van der Sluis (2003). When an object is not one of the objects that are in focus (i.e., –F) there is typically a larger set of distractors than when the object is in focus. As a result, reference to a –F object is more likely to involve a pointing act because the cost of the pointing act is balanced by the fact that with a large number of distractors a large number of linguistic properties is needed. In contrast, if the referent is +F, there are few distractors and consequently few properties are required to identify the object. In that situation, pointing is too expensive to replace verbal identification.

Finding 4: *Speakers tend to point more frequently when the referent is important. When the referent is important they point 55% (42/76) of the time, whereas when the referent is not important they point only 24% (11/45) of the time (highly significant at $P \leq 0.001$, $\chi^2 = 10.90$, df = 1). See*

[4] Cremers's (1994) notion of domain focus seems to bear some relation to the notion of an active object as proposed recently in Gergle (2006). Active objects are objects that were recently moved in the shared workspace. Roughly speaking, Cremers's notion of a domain focus can be viewed as often coinciding with the active object together with the objects in its immediate surroundings (in the current study, if an object was referred to in the preceding utterance, it was typically manipulated immediately after that, and would therefore count as an active object).

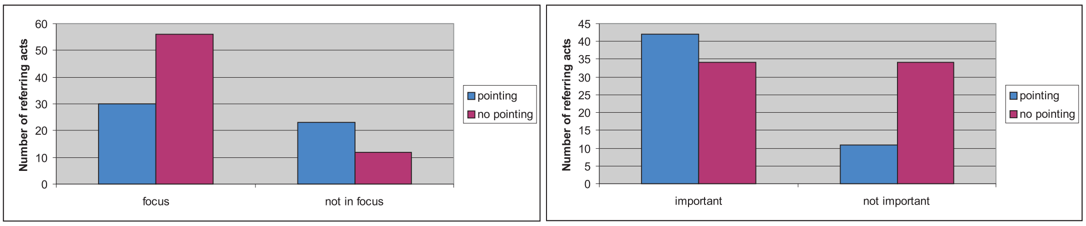


Figure 4: Number of initial referring acts with and without a pointing act mapped against whether the intended referent is in or out of domain focus, and important or not important.

Figure 4. For the purpose of the current study, an object was classified as + Important (+I) at time t if the speaker instructed the addressee to manipulate the object at t. All other objects, at the same point in time t, were labeled – Important (–I). Our finding on the influence of importance on modality choice does not seem to be reflected in any of the extant algorithms.

Finding 5: *Speakers appear to follow at least two different strategies when referring to objects. In particular, there seem to be two types of speaker: those that* never *use pointing and those that do use pointing (see Tables 1 and 2).* Approximately 29% of the speakers never use pointing (2I,12B,16I and 20I). For the remaining speakers, both −F and +I are good predictors for pointing: (a) for −F, 6 speakers point in the majority of cases, 2 speakers point in half of the cases (10I and 18B) and for the remaining 2 speakers we have no referring acts to −F objects; (b) for +I, 6 speakers point in the majority of cases, 2 do not (6I and 10I) and for the remaining 2 we have no referring acts to +I objects.

When we consider only the data for subjects that do at least point once, then we get the following statistics: (A) for referring acts to +F and –I objects, we have 8 referring acts involving pointing and 13 without pointing. (B) for referring acts to –F and +I objects, we have 20 referring acts involving pointing and 4 without pointing.

Participant		+F		–F	
Dialogue	Role	+P	–P	+P	–P
2	B	1	2	0	0
2	I	0	3	0	2
4	I	7	0	2	1
6	I	2	4	1	0
8	I	5	1	5	0
10	I	4	6	1	1
12	B	0	2	0	0
12	I	3	2	4	0
14	I	3	0	4	0
16	I	0	15	0	1
18	B	0	3	1	1
18	I	4	5	5	2
20	B	1	0	0	0
20	I	0	13	0	4

Table 1: Per dialogue participant (identified by dialogue number and role), the number of referring acts with and without pointing (+/– P) is listed for objects in and out of Focus (+/– F)

Participant		–I		+I	
Dialogue	Role	+P	–P	+P	–P
2	B	1	2	0	0
2	I	0	2	0	0
4	I	1	0	8	1
6	I	0	0	3	4
8	I	1	1	9	0
10	I	4	3	1	4
12	B	0	1	0	1
12	I	2	1	5	1
14	I	0	0	7	0
16	I	0	8	0	8
18	B	1	4	0	0
18	I	1	2	8	5
20	B	0	0	1	0
20	I	0	10	0	7

Table 2: Per dialogue participant (identified by dialogue number and role), the number of referring acts with and without pointing (+/– P) is listed for not Important and Important objects (–/+ I)

4 Strategies for Modality Choice

Based on the findings that are listed in the previous section we suggest that there are at least two (mutually exclusive) strategies for modality choice.

> Strategy 1: Never use pointing. Formulate the referring expression using only verbal means of expression.

> Strategy 2: Prefer pointing, if the intended referent is not in focus (–F) *or* important (+I).

Strategy 1 is directly based on the data from our observational study which had a substantial proportion of speakers that never pointed. For these speakers, the factors +/– F and +/– I consistently seemed to have no impact. We have not been able to find a factor that might explain why these speakers never pointed. Note that such a factor (or collection of factors) can take two very different forms: 1) there is something to the specific situations in which these speakers referred which made them refrain from pointing (and which would have made all the speakers involved in the study refrain from pointing). This would mean that if this factor is taken into account, we can merge the strategies 1 and 2 into a single strategy for all speakers. 2) The reason for not pointing lies not so much with the situation, but rather with the speakers themselves. This would mean that we have two types of speakers: those following strategy 1 and those following strategy 2. This is not so different from, for example, saying that there are speakers with different personalities (and consequently ways of expressing themselves) or, at the extreme end, individual styles. Here, we want to offer this second option as hypothesis that needs further investigation. The assumption that under the same circumstances different people have identical inclinations to refer by pointing is called into question by the current study and is in need of further investigation through controlled experiments. Additionally, note that observational reports from sociologist suggest that the inclination to point might vary between subcultures of the same language community (e.g., Scheflen, 1972).

Strategy 2 models speakers that do occasionally point. It covers in particular situations where the intended referent is –F or +I. We assume that speakers decide on pointing *before* they choose the verbal means that will accompany the pointing act.

The strategy 2 does not tells us what to do when the referent +F and –I. Our data, however, tell us that speakers most of the time do not point in under those circumstances. This suggests that in such situations speakers might very well apply an algorithm akin to the one advocated by

Krahmer and Theune (2003). Normally, this will lead to a referring act which contains no pointing, but if the number of distractors is sufficiently large, pointing might be included. Alternatively, it might be that also for the +F and –I situation there are other factors that *a priori* determine whether the speaker will point or not.

Note that strategy 2 can be seen as an efficient heuristic for approximating the outcomes of the Krahmer and van der Sluis (K&VdS) algorithm, particularly for those situations where the intended referent is –F. The efficiency derives from the fact that for –F referents, we expect the cheapest referring act to require a pointing act most of the time (because there will be a significant number of distractors as result of the fact that the set of distractors is not limited to those that are in domain focus), and therefore only considering referring acts that contain a pointing act – as proposed by strategy 2 – limits the search space.

In another respect, the K&VdS method does, however, differ from strategy 2: the K&VdS method does not take importance of the intended referent into account.

Finally, we would like to argue that there is a further consideration for the inclusion of pointing to –F referents (as proposed in strategy 2), that is not covered by the K&VdS method. Even if pointing is not the cheapest option in K&VdS's sense (when compared to verbal means of referring), the *attention directing* function of pointing (see, e.g., Butterworth, 2003:9) might lead to a preference for pointing. The use of pointing when the intended referent is –F, is more than simply a way of identifying the object, it also serves to *directly guide* the addressee's gaze to the referent. This is something which cannot be achieved by verbal means only (where the addressee will need to *interpret* the referring expression and then decide where to direct his or her gaze). In particular, if the addressee's gaze is in a region that is distant from the intended referent, pointing can be more adequate than verbal identification, because it allows the speaker to directly manipulate the addressee's gaze.[5]

5 Conclusions

In this paper, we have outlined our arguments against two predominant assumptions regarding modality selection in the generation of referring acts: the assumption that non-verbal means of referring are secondary to verbal ones, and the assumption that there is a single strategy that speakers follow for generating referring acts. To some extent, we have reversed these assumptions by arguing that (A) the choice regarding whether or not to point precedes the choice of verbal means of reference and (B) offering as a hypothesis for further investigation the claim that speakers might differ (irrespectively of situational factors) with regards to their inclination to refer by pointing.

Our supporting evidence was drawn from data obtained through an observational study. We specified two alternative strategies for modality selection based on correlation data from the observational study. Speakers that follow the first strategy simply abstain from pointing. Speakers that follow the other strategy make the decision whether to point dependent on whether the intended referent is in focus and/or important. Further controlled experiments are needed to verify whether the decision to point is effected by (rather than merely correlated with) whether the intended referent is in focus or important.

Let us conclude by identifying two further topics for research. These concern the precise realization of pointing acts and verbal expressions in multimodal referring acts. Firstly, on the basis of a case study of deictic gestures by Neapolitan speakers, Kendon and Versante (2003:134) suggest that 'the character of the pointing gesture itself might vary systematically in relation to semantic distinctions of various sorts'. Secondly, the choice of determiner also requires further study; most quantitative studies of the use of definite article, proximal and distal demonstrative concern their use in text, see e.g., Kirsner and Van Heuven (1988) and Maes and Noordman (1995). In our corpus, all these types of referring expression occur. Demonstratives of both the proximate and distal variety are abundant in our corpus (see Piwek and Cremers, 1996) and, interestingly, we also observed 11 instances where an indefinite noun phrase (e.g., 'Now there is, where you just were, a red square on the floor.') was used in a referring act.

[5] In other words, pointing is a form of indicating in the sense of C.S. Peirce; see Buchler (1940: chap. 7).

References

Bates, J. (1994). The role of emotion in believable agents. In: *Communications of the ACM* **37**(7), 122–125.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science* **15**(6), 415–419.

Beun, R.J. and Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition* **6**(1/2), 121–152.

Buchler, J. (1940). *The Philosophy of Peirce: Selected Writings.* Routledge and Kegan Paul, London.

Butterworth, G. (2003). Pointing is the Royal Road to Language for Babies. In: Kita, S. (Ed.), *Pointing: Where Language, Culture and Cognition Meet*, Lawrence Erlbaum, New Jersey, pp.9–34.

Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. (2000)(Eds.). *Embodied Conversational Agents*, The MIT Press, Cambridge, MA.

Claassen, W. (1992). Generating referring expressions in a multimodal environment, in: R. Dale et al. (eds.), *Aspects of Automated Natural Language Generation*, Springer Verlag, Berlin.

Clark, H. and Bangerter, A. (2004). Changing Ideas about Reference. In: I. Noveck & D. Sperber (Eds.), *Experimental Pragmatics* (pp. 25–49). Palgrave Macmillan, New York.

Cremers, A. (1993). Transcripties dialogen blokken-experiment. IPO report no. 889, Eindhoven.

Cremers, A. (1994). Referring in a shared workspace. In: Brouwer-Janse, M.D., Harrington, T.L., (Eds.), *Human-machine communication for educational systems design*, Springer Verlag, Heidelberg, 71–78.

Cremers, A. (1996). Reference to objects: an empirically based study of task-oriented dialogues. PhD thesis, Eindhoven University of Technology.

Dale, R. and E. Reiter (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions, *Cognitive Science* **18**:233-263.

Fitts, P. (1954). The information capacity of the human motor system in controlling amplitude of movement, *Journal of Experimental Psychology* **47**, 381-391.

Gergle, D. (2006). What's There to Talk About? A Multi-Modal Model of Referring Behavior in the Presence of Shared Visual Information. In: *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2006)* Student Research Workshop.

Grice, H.P. (1975). Logic and conversation. In: Cole, P. & J. Morgan (eds), *Syntax and Semantics 3: Speech Acts*, Academic Press, New York, 41–58.

Grosz, B. (1977). The representation and use of focus in dialogue understanding. *Technical note 151*, SRI International, Menlo Park.

Huls, C. and Bos, E. (1998). Studies into Full Integration of Language and Action. In: Bunt, H., Beun, R.J. & Borghuis, T. (1998). *Multimodal Human-Computer Communication; Systems, Techniques and Experiments*, Lecture Notes in Artificial Intelligence 1374, Springer, Berlin.

Kendon, A. and Versante, L. (2003). Pointing by hand in Neapolitan. In: S. Kita, (Ed.) *Pointing: Where Language Culture and Cognition Meet.* Lawrence Erlbaum, Hillsdale, N.J.

Khan, I., Ritchie, G., van Deemter, K. (2006). The clarity-brevity trade-off in generating referring expressions. In: *Proceedings of the Fourth International Natural Language Generation Conference (INLG)*, 14-15 July 2006. Sydney, Australia, 84-86.

Kirsner, R. and Van Heuven, V. (1988). The significance of demonstrative position in modern Dutch. *Lingua* **76**, 209–248.

Krahmer, E. and Van der Sluis, I. (2003), A New Model for Generating Multimodal Referring Expressions. In: *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003)*, April 12-13, Budapest Hungary, 47–54.

Kranstedt, A., Luecking, A., Pfeiffer, T., Rieser, H. and Staudacher, M. (2006). Measuring and Reconstructing Pointing in Visual Contexts. In: *Brandial 2006: The 10th Workshop on the Semantics and Pragmatics of Dialogue*, University of Potsdam, Germany, September 2006.

Kranstedt, A. and Wachsmuth, I. (2005). Incremental Generation of Multimodal Deixis Referring to Objects In: *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG 2005)*, Aberdeen, UK, August 2005, 75–82.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B. and Bhogal R. (1997). The persona effect: Affective impact of animated pedagogical agents, In: *Proceedings of CHI'97 (Human Factors in Computing Systems)*, 359–366, Atlanta.

Lester, J., J. Voerman, S. Towns and C. Callaway (1999). Deictic Believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents, *Applied Artificial Intelligence* **13**(4-5):383-414.

Maes, F. and Noordman, L. (1995). Demonstrative nominal anaphors: a case of nonidentificational markedness, *Linguistics* **33**, 255-282.

Neal, J. G., Thielman, C. Y., Dobes, Z., Haller, S. M., and Shapiro, S. C. (1989). Natural language with integrated deictic and graphic gestures. In: *Proceedings of the Workshop on Speech and Natural Language* (Cape Cod, Massachusetts, October 15 - 18, 1989). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 410–423

Piwek, P. and Cremers, A. (1996). Dutch and English Demonstratives: A Comparison. *Language Sciences*, **18**(3-4), 835-851.

Prendinger, H. and Ishizuka, M. (2004)(Eds). *Life-Like Characters: Tools, Affective Functions, and Applications*, Springer, Berlin.

Scheflen, A. (1972). *Body language and the social order: communication as behavioral control.* Prentice-Hall, Englewood Cliffs, N.J.

van der Sluis, I. and Krahmer, E. (2001). Generating Referring Expressions in a Multimodal Context: An empirically motivated approach. In: W. Daelemans et al. (eds.), *Selected Papers from the 11th CLIN Meeting*, Rodopi, Amsterdam.

Wilkins, D. (1999). Manual for the 1999 Field Season. Language and Cognition Group. Max Planck Institute for Psycholinguistics, Nijmegen.

Some Multimodal Signals in Humans

Jan Peter de Ruiter
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD NIJMEGEN, The Netherlands
janpeter.deruiter@mpi.nl

Abstract

In this paper, I will give an overview of some well-studied multimodal signals that humans produce while they communicate with other humans, and discuss the implications of those studies for HCI.

I will first discuss a conceptual framework that allows us to distinguish between *functional* and *sensory* modalities. This distinction is important, as there are multiple functional modalities using the same sensory modality (e.g., facial expression and eye-gaze in the visual modality). A second theoretically important issue is *redundancy*. Some signals appear to be redundant with a signal in another modality, whereas others give new information or even appear to give conflicting information (see e.g., the work of Susan Goldin-Meadows on speech accompanying gestures). I will argue that multimodal signals are never truly redundant. First, many gestures that appear at first sight to express the same meaning as the accompanying speech generally provide extra (analog) information about manner, path, etc. Second, the simple fact that the same information is expressed in more than one modality is itself a communicative signal.

Armed with this conceptual background, I will then proceed to give an overview of some multimodal signals that have been investigated in human-human research, and the level of understanding we have of the meaning of those signals. The latter issue is especially important for potential implementations of these signals in artificial agents.

First, I will discuss pointing gestures. I will address the issue of the *timing* of pointing gestures relative to the speech it is supposed to support, the *mutual dependency* between pointing gestures and speech, and discuss the existence of alternative ways of pointing from other cultures. The most frequent form of pointing that does not involve the index finger is a cultural practice called lip-pointing which employs two visual functional modalities, mouth-shape and eye-gaze, simultaneously for pointing.

Next, I will address the issue of eye-gaze. A classical study by Kendon (1967) claims that there is a systematic relationship between eye-gaze (at the interlocutor) and turn-taking states. Research at our institute has shown that this relationship is weaker than has often been assumed. If the dialogue setting contains a visible object that is relevant to the dialogue (e.g., a map), the rate of eye-gaze-at-other drops dramatically and its relationship to turn taking disappears completely. The implications for machine generated eye-gaze are discussed.

Finally, I will explore a theoretical debate regarding spontaneous gestures. It has often been claimed that the class of gestures that is called *iconic* by McNeill (1992) are a "window into the mind". That is, they are claimed to give the researcher (or even the interlocutor) a direct view into the speaker's thought, without being obscured by the complex transformation that take place when transforming a thought into a verbal utterance. I will argue that this is an illusion. Gestures can be shown to be specifically designed such that the listener can be expected to interpret them. Although the transformations carried out to express a thought in gesture are indeed (partly) *different* from the corresponding transformations for speech, they are a) complex, and b) severely understudied. This obviously has consequences both for the gesture research agenda, and for the generation of iconic gestures by machines.

Keywords: Natural Interaction, gesture, eye-gaze, multimodality.

1 MODALITIES AND REDUNDANCY IN HUMAN-HUMAN COMMUNICATION

The word “multimodal” refers to “multiple modalities”, but the word “modality” is ambiguous. Humans have five sensory modalities, corresponding to our five senses: vision, hearing, haptic perception, smell, and taste. The human organism has specialized receptors for each of these sensory modalities. However, in the field of multimodal communication (either between humans or between humans and artificial agents), a different sense of the word “modality” is used. For instance, eye-gaze and gesture are usually seen as distinct modalities, even though they are both -- technically speaking -- visual modalities, as the perception of both eye-gaze and gesture both rely on visual perception. So what then is a modality in the context of multimodal communication, and how can we define it independently of the sensory modality that is involved? De Ruiter, Rossignol, Vuurpijl, Cunningham & Levelt (2003) have proposed to make a distinction between *sensory* modalities on the one hand, and *functional* modalities (called *semiotic channels*) on the other. Sensory modalities are easily defined: they correspond to the sensory system that is involved. Functional modalities, however, are independent of sensory modalities, although there is always a particular sensory modality associated with a functional modality. To distinguish between multiple functional modalities that have the same underlying sensory modality, De Ruiter et al. have proposed to use the following criteria. Within a functional modality, all the identifiable signals that can be produced in that modality should be mutually exclusive. That is, within one functional modality it is not possible to produce two different signals at the same time. However, between functional modalities, every signal in modality A should be combinable with any other signal in modality B. For example, voice quality and facial expression are both considered functional modalities, as a) it is not possible to have two voice qualities at the same time, b) it is not possible to have two facial expressions at the same time, and c) it is in principle possible to combine any voice quality with any facial expression. Using this approach, we can also motivate why gesture and facial expression are different functional modalities, even though they rely on the same perceptual modality. An interesting case that seems to defy this system is presented by eye-gaze and facial expression. Although they are normally independent of each other, there are specific facial expressions that rely on a specific eye-gaze movement, for instance the “rolling eyes” expression to indicate disapproval. Although the proposed system is not entirely waterproof, it is useful in distinguishing modalities such as prosody, intonation, voice quality and speech content, all relying on the auditory modality. Similarly, one can distinguish eye-gaze, facial expression, head-tilts, and hand-gestures even though they are all relying on the visual modality.

Now that we have a motivated working definition of modalities in multimodal communication, it is interesting to look at the phenomenon of redundancy between modalities. Goldin-Meadow and colleagues (Alibali & Goldin-Meadow, 1993, Goldin-Meadow, Alibali, & Church, 1993) distinguish between two types of gestures: gestures that either *match* or *mismatch* the concept expressed in the accompanying speech. If a child that is reasoning about Piaget's conservation task would describe a container using the word “tall” while gesturing the concept of “skinny” by putting the hands vertically close to each other, this would be a mismatch, as the gesture expresses another concept than the speech. If, on the other hand, the child would say “skinny” while making the same gesture, this would count as a “match”. While the distinction that these authors make appears to be useful for their experimental purposes, it is important to realize that iconic gestures are never truly redundant with their speech: they always provide information that is not in the speech (Kendon, 2004, De Ruiter, in press). In the example above, the gesture for “skinny” did not only indicate the concept of skinniness, but also *how* skinny the particular container was. In gestures that depict motion (e.g., “falling”) the gesture contains information about the speed and path of the falling that is being described. While this holds for iconic gestures, it appears that for *emblems*, gestures that are culture-specific idiosyncratic equivalents of certain verbal concepts (the gesture for “three”, or “OK” in English for instance) that they can truly be redundant. However, saying “OK” with the “OK” gesture is still, from a communicative point of view, different from saying “OK” without the gesture, simply because the gesture is added (or not). Although from a formal point of view, adding a gesture can be considered redundant or resulting in *overspecification* (Van der Sluis, 2005), from a social perspective the added redundancy is itself a signal. Careful qualitative analysis of natural conversation reveals that whether a potentially redundant gesture is produced or not alters the interpretation of the listener, and speakers are aware of whether it is deemed redundant or not. Enfield, Kita and De Ruiter (submitted) have found that if people point at locations while suspecting that their interlocutor might well be aware of the mentioned location (hence risking to provide their interlocutor with information that they already have) they make much smaller and more inconspicuous pointing gestures than when they suspect their interlocutor is not aware of the mentioned location. Producing a big and salient pointing gesture to a known location is not only redundant, it is also potentially a threat to the interlocutor's face.

2 POINTING: WHEN, WHY, AND HOW?

Pointing gestures appear on first sight to be relatively straightforward to understand. Popular conceptions are that they occur obligatorily with deictic expressions such as "this" or "that" (Levelt, Richardson, & La Heij, 1985, De Ruiter & Wilkins, 1998), but that pointing gestures can be understood without speech (Tomasello, Carpenter, Call, Behne, & Moll, 2005), and that the canonical pointing gesture involves the use of the index finger. In the following, I will report on some findings from human-human communication that may shed a different light on these conceptions.

2.1 POINTING: WHEN?

It is often assumed that pointing gestures occur simultaneously with the deictic term they are accompanied by. In fact, McNeill (1992) has claimed that due to the common origin of gesture and speech, the gesture and speech that originate from the same growth point are produced simultaneously, as gesture and speech are in fact *simultaneous* expressions of the same underlying idea (called a *Growth Point*). This assumption has been implemented in virtual agents like REA (Cassell et al., 1999) in which both the beginning and the end of the generated gesture are made to match with the beginning and end of the accompanying speech. Making both the beginning and end of a gesture and speech interval coincide in an agent is very difficult, especially when it comes to making the *end* of the intervals co-occur (Cassell, personal communication). But perhaps it is not necessary to synchronize the end of gestures and the accompanying speech at all.

An alternative proposal about the relationship between gesture and speech is that although gesture and speech are planned together, their actual execution is "ballistic", in the sense that after a shared planning stage, the execution of the gesture and the speech proceed without any interprocess communication (Levelt et al., 1985, De Ruiter, 1998, De Ruiter & Wilkins, 1998). If this is actually how humans point and speak, the synchronization of gesture and speech can be implemented much more loosely than in agents like REA.

This discussion is somewhat complicated by two definitional problems. First, it is not always clear what the part of speech is that the gesture corresponds to. Second, the movement pattern of hand gestures consist of several phases (Kendon, 1980). First, there is a preparation phase, in which the hand moves from a resting position (or the final position resulting from a previous gesture) to the beginning point of the next phase, the so-called *stroke*. It is the stroke that carries the primary communicative information. Then, after completing the stroke, there is a retraction phase in which the hand(s) return from the final position of the stroke to a resting position. With pointing gestures however, the most meaningful part is what Levelt et al. (1985) have called the *apex*, the moment in time that the outstretched hand performing a pointing gesture is held still for a short period of time. This notion is in conflict with Kendon's (1980) notion that the stroke corresponds with the part of the gesture with maximal effort. So for pointing gestures, it makes sense to assume that the stroke of the gesture is the interval between the beginning of the (directed part of the) motion and the apex.

There is ample evidence that the onset of pointing gestures actually *precedes* their spoken affiliates by a fraction of a second (Levelt et al., 1985, De Ruiter & Wilkins, 1998). This finding has later been confirmed in work by Foster (Foster, 2004). In De Ruiter (1998), the initiation of pointing gestures that are produced simultaneously with definite referring expressions (without deictics) was shown to also start more than 200 ms. before the onset of the first word of the referring expression (in this case the definite article). For multimodal output generation, this means that for the most naturalistic results, pointing gestures should be initiated before the related speech is produced. Kopp & Wachsmuth (2004) have taken the general architecture presented in De Ruiter (2000) to implement an artificial agent that displays this timing behavior.

2.2 POINTING: WHY?

In a project on the role of dialogue in Joint Action, we have closely observed dyads who construct toy models together from component parts that were placed on a table in between the subjects. Dyads were recorded by three synchronized high resolution video cameras. In half of the conditions, the subjects were allowed to speak, and in the other half they were not. In the condition without speech, we expected to see many stand-alone pointing gestures, for instance when subjects want to direct their interlocutor's attention to a specific object without the use of speech. We analyzed the different types of actions that subjects performed, such as *put-down*, *grasp*, *release*, *give*, *point*, etc.

Interestingly, the subjects who could speak produced a lot more pointing gestures than the subjects who didn't speak. In the speaking condition, 10.3 % of all recorded actions were pointing gestures, whereas in the silent condition only 2.1 % were pointing gestures. From this finding we concluded that speech and pointing are mutually dependent on one another. It was already known from previous work (e.g., Levelt et al., 1985) that there are parts of speech such as deictic expressions that require some form of gesture to go along with them, but apparently the reverse holds too: pointing gesture require speech to go along with the gesture. While it is possible and perhaps even functional to point without speaking, in the silent condition our subjects chose other means of indicating objects, and not pointing.

In generating pointing gestures in artificial agents therefore, using stand-alone pointing gestures without accompanying speech will require a careful pragmatic analysis to find out how such stand-alone pointing gestures are interpreted by humans. It is probably beneficial for the naturalism of the behavior of the artificial agent to have it point a) together with some form or referring expression, and b) have the pointing gesture start before the accompanying referring expression is produced.

2.3 POINTING: HOW?

It is commonly believed that the natural way for humans to point is to create a vector in 3D space with our outstretched index finger (Morris, 1978, Povinelli & Davis, 1994). This turns out not to be a universal. Wilkins (2003) argues on the basis of an extensive inspection of the pointing systems that are in use by many different cultures, that a) there are many cultures, to be found on all inhabited continents, whose dominant way of pointing is not with the index finger, but by lip-pointing[1], and b) in cultures that also have these other forms of pointing, the "standard" index finger point is often still used as well, but has another meaning (usage) than our western index finger pointing. These two observation lead Wilkins to conclude that pointing with the index finger is a cultural practice that is learned/acquired, and not a universal. In multimodal output generation for artificial agents, it is of course possible to restrict oneself to the index finger pointing that is predominant in our western culture, but this restriction may well represent a missed opportunity to understand multimodal referring in its full glory. In Arrernte, the Australian Aboriginal culture studied extensively by Wilkins (2003), there are, in addition to lip pointing, several different hand shapes that distinguish (among other things) between pointing to an object, an area or set of objects, a direction, or a path. Furthermore, in referring to non-visible objects, the vertical angle of the arm during pointing can indicate whether the referent is in proximal, mid-distance, or distal space.

But we do not have to travel to the other side of the planet to be confronted with different types of pointing. In Neapolitan Italian, index finger pointing with the palm downwards is used in singling out objects that are the primary focus of the discourse, whereas pointing with the palm vertical is used to indicate objects that have relevance for the discourse, but are not themselves the focus (Kendon & Versante, 2003).

Artificial Agents that point would ideally have architectures in which these cross-cultural variations could be specified or parameterized.

[1] Interestingly, Enfield (2001) showed that lip-pointing is not the same as pointing with the lips. In fact, according to Enfield, in lip-pointing it is the eye-gaze that is most accurately pointing at the intended location, whereas the protruded lips function to indicate that the eye-gaze is now doing the pointing. Here we see two functional modalities collaborate to perform one specific communicative function.

3 EYE-GAZE AND CONVERSATION DURING VISUAL TASKS

In a classic study by Kendon (1967), a number of observations are made about the relationship between turn-taking and eye-gaze-at-the-interlocutor (henceforth called "other-gaze") that are very influential, and have been implemented in artificial agents. Kendon's main observations, replicated by Torres, Cassell, & Prevost (1997), are that other-gaze is predominantly a *listener* phenomenon. Slightly simplified and idealized, the basic idea is this: listeners do other-gaze, speakers do so much less, but towards the end of their conversational turn speakers will initiate a new interval of other-gaze, indicating that their turn is about to end and they are about to take the listener role again.

It is important to note that these studies have looked at eye-gaze in face to face conversation, and that these conversations were not about something that was visible in the same space. This is important, because Argyle & Graham (1976) found that as soon as there is an object that is relevant for the discourse (e.g., a map or a picture that is the topic of the conversation) present in the shared visual field of both interlocutors, there is a dramatic reduction of other-gaze. While in ordinary conversation the rate of other-gaze is about 60-75%, in the Argyle and Graham experiments it dropped to less than 10%. If such a relevant object is present, the participants in the dialogue will direct their eye-gaze towards this object. There may be multiple reasons for this phenomenon. One interpretation is that the relevant visual object needs to be looked at for information processing reasons. Another, alternative explanation is that other-gaze signals aggression in many higher mammals. In other words, looking at the interlocutor might well be uncomfortable. However, the rules of politeness dictate that it is not acceptable for listeners to look out of the window while their interlocutor is speaking, as this may be interpreted as a lack of interest. In the case of a mutually visible object that is relevant to the dialogue, looking at that object instead of the interlocutor poses no threat to the face of the interlocutor, as looking at that object appears to be directly related to the conversation.

The relevance of the findings of Argyle and Graham in the context of multimodal output are that in HCI, but also in many human-human communication studies, there is a certain activity that is performed by the human and/or computer. There usually is a task. And the task usually involves visual representations that are relevant. If this is the case, what should we do with the eye-gaze of our artificial agent?

To address this question, it is important to know whether the low frequency other-gaze in visual tasks still has the relation with turn-taking in conversation. In the context of the COMIC (Conversational Multimodal Interaction with Computers) project, we have done a study to address this issue (De Ruiter, 2005). Four dyads were given sets of picture pairs that were subtly different from one another. The participants were asked to find out through dialogue what the difference was between their pictures, without looking at the picture of the other. This task situation generated lively dialogues, and (as expected) most of the time the eye-gaze was directed at the picture. The proportional duration of eye-gaze directed at the interlocutor (by either participant) was only 7%. The main question now was whether this low frequency other-gaze was still coupled with the turn-taking. Using a contingency analysis of the combined turn-taking and eye-gaze states (see De Ruiter 2005 for details) it was established that there was no relationship at all between the turn taking state (who is/are speaking) and the gaze-state (who is/are looking at the other). In addition, there were no differences in other-gaze frequency on a number of factors that were analyzed. These factors were a) whether the speakers had just finished either a question or a statement, b) whether the object referred to was easy or hard to refer to[2], and c) whether the object referred to was mentioned for the first time, or for the N-th (N>1) time. We analyzed these factors because we suspected that there would be more other-gaze after hard or initial references (to check whether the interlocutor got it) and after questions (to check for possible nonverbal responses). But for all these factors, the frequency of other-gaze were statistically not distinguishable.

In a collaborative project with Douglas Cunningham at the Max Planck Institute for Biological Cybernetics in Tübingen, we implemented an other-gaze control module for the conversational agent of the COMIC demonstrator that incorporated these findings by generating randomized other-gaze with the same frequency distribution as the humans from the study mentioned above. As the COMIC agent was

[2] The materials were constructed such that half of the depicted objects could not be described with a monomorphemic noun, meaning the subjects had more trouble referring to these hard-to-name objects.

discussing bathroom design with customers, while a CAD application was displaying various expensive bathrooms, we suspected that low-frequency random other-gaze would be quite appropriate in this context. Although no formal user study has been done yet, informal tests with the agent indicated that the low frequency random other-gaze has a very naturalistic effect. Subjectively, it appears that the agent is engrossed in the (shared) visual workspace, but checks every now and then whether its interlocutor is still there. Further studies on this topic will need to shed more light on the issue of other-gaze frequency in agents.

4 THE GENERATION OF ICONIC GESTURES; A VERY HARD PROBLEM

In empirical gesture research, many studies have been done on what McNeill (1992) has termed *iconic* gestures. Iconic gestures have a number of interesting properties. First, their shape resembles their referent, which is the sense in which they are "iconic". Furthermore, iconic gestures are not lexicalized the way *emblematic* (McNeill, 1992) or *quotable* (Kendon, 1990) gestures[3] are. While quotable gestures have a form/meaning mapping that is arbitrary and shared within a linguistic community, iconic gestures are generated on the fly by speakers, and have no pre-defined form/meaning mapping.

The fact that iconic gestures resemble their referents and are not governed by shared and arbitrary form/meaning mappings has led several gesture researchers to claim that gestures are a "window into the mind" (Beattie, 2003, Goldin-Meadow et al., 1993, McNeill, 2000), which means that iconic gestures give the researcher and perhaps even the interlocutor privileged access to the "raw" representations that are active in the speaker's mind during communication. Whereas we know that to transform a communicative intention into audible speech involves many transformations before the actual signal is realized, the Window Into the Mind (WIM) hypothesis suggests that these transformations are not involved in the generation of iconic gestures.

As I have argued in De Ruiter (in press), iconic gestures also, like speech, undergo a series of transformations that, albeit different from those involved in speech, are sufficiently complex to prevent us from peeking into the mind directly through iconic gestures. In realizing the gestural part of a communicative intention, at least the following transformations are necessary to end up with an observable gesture.

First, there is a process of *information selection*. Not all information that is available to the speaker is expressed in gesture. There is a selection process that determines what information is going to be expressed in gesture, and this partly depends on which information is selected to be expressed in speech (see e.g., Melinger & Levelt, 2004). Second, there is the assignment of a perspective (McNeill, 1992). Gestures can take a protagonist perspective (often leading to the subclass of *pantomimic* gestures), or an outside observer perspective, which can also vary. Third, and perhaps most importantly, iconic gestures are designed such that the interlocutor can (together with the accompanying speech) be expected to extract the intended meaning from them. The fact that listeners (and gesture researchers) find it relatively easy and straightforward to extract the communicative intention from an iconic gesture (again, with the help of the accompanying speech) does not come for free. In other words, iconic gestures are subject to *recipient design* (Sacks & Schegloff, 1979) also called *audience design* (Clark & Carlson, 1982). Ironically, the fact that we can effortlessly interpret the meaning of iconic gestures may have lead researchers to believe they are a window into the mind, but to accomplish this ease of interpretation the gestures need to be carefully designed, which rules out the possibility that they are windows into the mind.

Additional evidence for the claim that iconic gestures have to be designed is provided by those who build artificial agents that gesture. Although there are sophisticated agents that gesture quite convincingly (see e.g., Kopp, 2003, Kopp & Wachsmuth, 2004), there is to date no gesturing agent that is able to generate spontaneous iconic gestures on the fly from those spatial representations that happen to be active in its short term memory. This is understandable, as the computational problems are daunting. In fact, we have not even begun to understand how humans do this. To gesture convincingly, the gesturing agent should perform all the transformations mentioned above to a satisfactory degree, and then generate a motor program from it. Although progress has been made with respect to the generation of realistically

[3] This could also be the case for pointing gestures, see the discussion in section 2.3 above.

timed and executed motor programs (as witnessed by the work by Kopp and Wachsmuth quoted above), the spontaneous generation of iconic gestures on the basis of the communicative intention has so far eluded us. I therefore wish to end this paper by encouraging researchers in the field of multimodal output generation to spend more resources on this fascinating aspect of multimodal communication.

REFERENCES

Alibali, Martha, & Goldin-Meadow, Susan. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468-523.

Argyle, Michael, & Graham, Jean A. (1976). The central Europe experiment: Looking at persons and looking at objects. *Environmental Psychology & Nonverbal Behavior, 1*(1), 6-16.

Beattie, Geoffrey. (2003). *Visible Thought: The New Psychology of Body Language*. Hove, GB.: Routledge.

Cassell, J., Bickmore, T., Billinghurst, L., Campbell, K., Chang, H., Vilhjalmsson, H., et al. (1999). Embodiment in conversational faces: Rea. *Proceeding of CHI' 99*, Pittsburgh, PA., pp. 520-527.

Clark, Herbert H., & Carlson, Thomas B. (1982). Hearers and speech acts. *Language* (58), 332-373.

Enfield, N.J. (2001). 'Lip-pointing' - a discussion of form and function with reference to data from Laos. *Gesture, 1*(2), 185-212.

Foster, Mary Ellen. (2004). Corpus-based planning of deictic gestures in COMIC. Paper presented at the *Third International Conference on Natural Language Generation*.

Goldin-Meadow, Susan, Alibali, Martha, & Church, R. Breckinridge. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review, 100*, 279-297.

Kendon, Adam. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22-63.

Kendon, Adam. (1980). Gesticulation and speech: Two aspects of the process of an utterance. In M. R. Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207-227). The Hague: Mouton.

Kendon, Adam. (1990). Gesticulation, quotable gestures and signs. In M. Moerman & M. Nomura (Eds.), *Culture Embodied.* (Vol. 27, pp. 53-77). Osaka: National Museum of Ethnology.

Kendon, Adam. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kendon, Adam, & Versante, Laura. (2003). Pointing by hand in 'Neapolitan'. In S. Kita (Ed.), *Pointing: Where Language, Culture and Cognition Meet.* (pp. 109-137). Hillsdale, N.J.: Lawrence Erlbaum.

Kopp, Stefan. (2003). *Synthese und Koordination von Sprache und Gestik für virtuelle multimodale Agenten*. Berlin: Akademische Verlagsgesellschaft Aka GmbH.

Kopp, Stefan, & Wachsmuth, Ipke. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds, 15*, 39-52.

Levelt, Willem J.M., Richardson, Graham, & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language, 24*, 133-164.

McNeill, David. (1992). *Hand and Mind.* Chicago, London: The Chicago University Press.

McNeill, David (Ed.). (2000). *Language and Gesture: Window into Thought and Action*. Cambridge: Cambridge University Press.

Melinger, Alissa, & Levelt, Willem J.M. (2004). Gesture and the communicative intention of the speaker. *Gesture, 4*(2), 119-141.

Morris, D. (1978). *Manwatching: A Field Guide to Human Behavior.* St. Albans, England: Triad Panther.

Povinelli, D.J., & Davis, D.R. (1994). Differences between chimpanzees (*Pan troglodytes*) and humans (*Homo Sapiens*) in the resting state of the index finger: Implications for pointing. *Journal of Comparative Psychology, 108*, 134-139.

De Ruiter, J.P. (1998). *Gesture and Speech Production.* Unpublished Dissertation, Nijmegen, Nijmegen.

De Ruiter, J.P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture* (pp. 284-311). Cambridge, UK: Cambridge University Press.

De Ruiter, J.P. (2005). If eye-gaze frequency drops, its relationship with turn-taking disappears. Poster presented at *AMLAP 2005*, Ghent, Belgium.

De Ruiter, J.P. (in press). Postcards from the mind: the relationship between thought, imagistic gesture, and speech. *Gesture*.

De Ruiter, J.P., Rossignol, Stéphane, Vuurpijl, Louis, Cunningham, Douglas C., & Levelt, Willem J.M. (2003). SLOT: A research platform for investigating multimodal communication. *Behavior Research Methods, Instruments, & Computers, 35*(3), 408-419.

De Ruiter, J.P., & Wilkins, David P. (1998). The synchronisation of gesture and speech in Dutch and Arrernte (an Australian Aboriginal language): a cross cultural comparison. Paper presented at the *Conférence Oralité et Gestualité*, Besançon, France.

Sacks, Harvey, & Schegloff, Emmanuel A. (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (Ed.), *Everyday Language: Studies in Ethnomethodology* (pp. 15-21). New York: Irvington.

Tomasello, Michael, Carpenter, Malinda, Call, Josep, Behne, Tany, & Moll, Henrike. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675-691.

Torres, O. E., Cassell, J., & Prevost, S. (1997). Modeling gaze behavior as a function of discourse structure. Paper presented at the *First International Workshop on Human-Computer Conversations*, Bellagio, Italy.

Van der Sluis, Ielka. (2005). *Multimodal Reference.* Unpublished Doctoral Dissertation, Katholieke Universiteit Brabant, Tilburg.

Wilkins, David P. (2003). Why pointing with the index finger is not a universal (in sociocultural and semiotic terms). In S. Kita (Ed.), *Pointing: Where Language, Culture and Cognition Meet* (pp. 171-215). Hillsdale, NJ: Lawrence Erlbaum.

Summarizing Dive Computer Data: A Case Study in Integrating Textual and Graphical Presentations of Numerical Data

Somayajulu G. Sripada and Feng Gao
Department of Computing Science
University of Aberdeen, Aberdeen, UK
{ssripada,fgao}@csd.abdn.ac.uk

Abstract

Currently numerical data are most often presented graphically to human users. A recent study in the medical domain showed that under experimental conditions medical staff made better clinical decisions when presented with human written textual descriptions of patient data than when they viewed graphical presentations of the same data. In another recent study in the domain of marine weather forecasts, regular readers of these forecasts rated computer generated texts higher than human written texts. These studies suggest that textual presentations are at least as good as graphical presentations and NLG can be used to produce high quality texts. All this means time is ripe for integrating textual and graphical presentations to help users get most out of their data. In this paper, we describe a case study on presenting dive computer data to scuba divers integrating text and graphics. Our integration of text and graphics is based on the need to ground phrases from the text describing patterns detected in the dive profile data which are marked up on the line graph showing the dive profile data. We also describe an evaluation study we carried out with scuba divers. Results from the evaluation study indicate that divers find bi-modal (text+graph) presentations more useful than uni-modal ones to judge the safety of a dive. Also divers find annotations on the line graph useful for interpreting the associated textual description.

Keywords: Natural language generation, scuba diving, time series data analysis, bi-modal (text+graph) output.

1 Introduction

In several domains, humans are required to understand large quantities of numerical data; in particular time series data, a series of data values corresponding to different time points. Medical staff in intensive care units (ICUs) inspect time series of physiological data in order to monitor patient health (Hunter et al., 2003). Maintenance engineers in charge of an operational gas turbine examine time series of sensor data in order to monitor the health of the turbine (Yu et al., 2006). In these domains, raw data needs to be summarized to improve data accessibility to human users. From earlier work on data summarization we can identify two tasks:

1. Information extraction – to separate required information from unwanted noise. There have been a number of techniques developed in the data mining community to extract 'significant' or 'novel' or 'surprising' portions of information from raw input data (Dasgupta and Forrest, 1996; Keogh et al., 2001; Keogh et al., 2002; Lin et al., 2002; Keogh et al. , 2005).

2. Information presentation – to present information in a way suitable to the user. Presentations of quantitative information are most often graphical (Tufte, 2001; Shneiderman and Plaisant, 2005). In the field of information visualization (InfoVis), several techniques are developed for presenting quantitative information using computer graphics (Shneiderman, 1996; Plaisant et al., 1998; Weber et al., 2001).

There have been several studies integrating the above two tasks of information extraction and information presentation to improve data accessibility to users (Fayyad et al., 2002). These studies have particularly focused on integrating data mining and information visualization. In the field of natural language generation (NLG) too there have been several attempts at producing textual summaries of quantitative information (Boyd, 1997; Sripada et al. 2003a; Sripada et al., 2003b; Yu et al., 2006) integrating data mining and NLG. However, compared to the number of studies integrating visual presentations to data analysis, the number of studies integrating textual presentations to data analysis has been low.

A recent study in the medical domain showed that under experimental conditions medical staff made better clinical decisions when presented with human written textual descriptions of patient data than when they viewed graphical presentations of the same data (Law et al., 2005). A similar result in another domain is reported in (Langan-fox et al., 2006). In another recent study in the domain of marine weather forecasts, regular readers of these forecasts rated computer generated texts higher than human written texts (Reiter et al., 2005). These studies suggest that textual presentations are at least as good as graphical presentations and NLG can be used to produce high quality texts. All this means time is ripe for integrating textual and graphical presentations to help users get most out of their data. Multimodal presentations of information have been studied earlier addressing issues such as modality choice and information distribution among the different modalities (Andre, 2000; Theune, 2001). PostGraphe is a system that produces a summary report of statistical data integrating text and graphics (Fasciano and Lapalme, 1996). CogenTex's ChartExplainer also produces reports that show statistical graphics and text. Although these studies integrate text and graphics to report quantitative information these studies do not explore the semantic dependencies between text and graphics that we study in this paper.

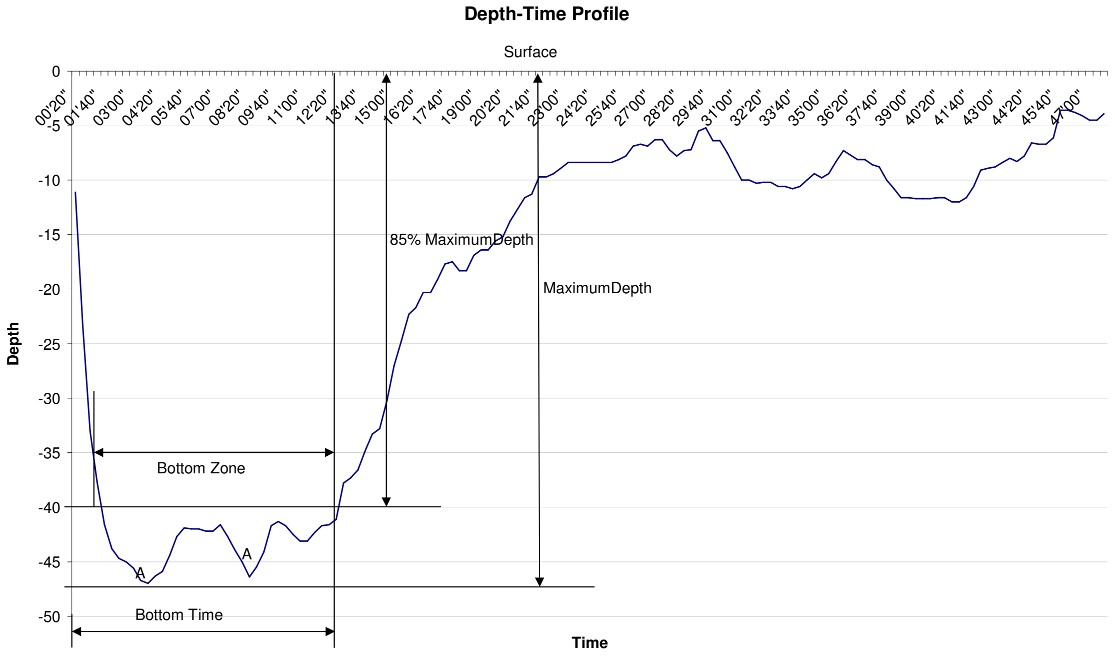


Figure 1: Depth-time profile of an example dive.

Risky dive with some minor problems. Because your bottom time of 12.0min exceeds no-stop limit by 4.0min this dive is risky. But you performed the ascent well. Your buoyancy control in the bottom zone was poor as indicated by 'saw tooth' patterns marked 'A' on the depth-time profile.

Figure 2: Textual summary of safety related information produced by ScubaText prototype for the example dive shown in Figure 1.

In this paper, we present a case study on presenting dive computer data to scuba divers. In this study, we developed a method for presenting safety related information of an archived dive data set using the combination of a textual summary and an annotated line graph. The textual summary of the report is based on interpretation of different structural elements (local patterns and global shape) of the depth-time profile of a dive. There are no universally accepted definitions of terminology used to refer to the structural elements of a dive profile. Therefore, when the structural elements of a dive profile are mentioned in a textual summary they are marked visually in the associated graphic of the depth-time profile. In other words, we use the graphic to provide grounding for our usage of the terminology. Verbal descriptions of the structural elements can be included in the textual summary to eliminate the need for a graphic completely. But these descriptions may distract the user from the safety related messages communicated in the text. We will also describe an evaluation study we carried out with scuba divers. Results from the evaluation study indicate that divers find bi-modal (text+graph) presentations more useful than uni-modal ones to judge the safety of a dive. Also divers find annotations on the line graph useful for interpreting the textual description.

Figure 1 shows a line graph representing the depth-time profile of a dive captured by a dive computer. Time is plotted along the X-axis and depth is plotted along the Y-axis. The structural elements such as 'bottom zone' and 'saw tooth' patterns are marked as annotations on the line graph thus providing grounding for their usage in the textual summary generated by the ScubaText prototype (described later) shown in Figure 2.

2 Background

Dive computers are electronic devices that guide SCUBA (Self Contained Underwater Breathing Apparatus) divers underwater in real time by providing instructions about safe bottom times and safe ascent schedules based on dive data such as depth-time profile captured using sensors (see http://en.wikipedia.org/wiki/Dive_computer). It's long been known that badly performed scuba dives cause decompression illness (DCI), a term used to refer to disease(s) caused by reduction in ambient pressure. Professional scuba divers and scientific scuba divers perform well supervised dives that strictly adhere to standardized safety regimes. However, recreational scuba divers who do receive training to perform safe dives during their certification courses, dive subsequently pretty much on their own and require support to cultivate safe diving habits.

The data recorded by a dive computer can be uploaded to a PC and can be explored using the vendor supplied software tool which presents the dive data visually using multiple coordinated views (MCV) (Baldonado et al., 2000). While the real time feedback from a dive computer is useful to divers in performing a particular dive safely, the vendor supplied visualization tool helps divers to assess their past dives in their leisure and learn important lessons about their diving behavior in general.

The vendor supplied visualization tool enables users to perform exploratory data analysis (EDA) (Tukey, 1977) of dive data. For example, the tool requires users to draw insights into the dive data by inspecting excess gas loadings in different body tissues displayed in one view (histogram view) in relation to specific observable patterns on the depth-time profile displayed in another view (line graph view). EDA may not be suitable to all divers. Some lack the time EDA requires and others may find EDA stressful because it requires focused use of mental skills to interpret multiple graphical views of numerical data. Improving access to dive data by other means is necessary.

2.1 SCUBATEXT

ScubaText is a research project that aims to produce feedback reports of scuba dives based on data recorded by a dive computer. Patterns on the depth-time profile of a dive are known to be helpful in judging the safety of the dive. Depth-time profile data can be modelled as time series, series of depth values recorded at regular intervals by the pressure sensor on the dive computer. ScubaText builds on work on summarising time series data carried out in a previous project, SumTime (Reiter et al., 2005a). Initially we tried to directly apply SumTime technology to summarise dive computer data. One important observation we made from this initial study was that summaries produced by SumTime technology were more descriptive while summaries required in the ScubaText were more interpretative, summaries that describe interpretations of important information extracted from raw data. In the ScubaText project we use domain knowledge to interpret patterns discovered from raw depth-time profile data. Interpreting the input data and describing the interpretations in the output summaries is one of the extensions ScubaText makes to the techniques developed in SumTime.

Initially using an earlier version of the ScubaText prototype, we performed a small study with dive instructors working at a local diving school, Aberdeen Water Sports. In this study we showed them a textual summary (different from the current version) and its associated depth-time profile graph without any annotations and asked them to comment qualitatively about the accuracy of the textual descriptions in relation to the accompanying depth-time profile. They made several comments on the content of the summaries which led us to the current version. But while speaking to them we realised the variations in the usage of technical terms in this domain. This is not the individual variation of usage (also known as idiolect) but the variations due to the affiliations to different professional bodies involved in diver training and the different vendors of dive computers. In the ScubaText project we use annotations on the dive profile graph to ground the meanings of these non-standard terms. Using annotations on graphical presentations for semantic grounding of certain phrases is another extension to the techniques developed in SumTime. In the SumTime-Turbine system one of the tasks is to choose words or phrases to refer to patterns in the sensor data when domain experts fail to agree to a single consensus word or phrase (Reiter et al., 2005a). Annotations marked on time series sensor data plots can help to communicate meaningful texts in the gas turbine domain. We believe that the idea of using annotations on graphs to ground non-standardised words or phrases referring to observable patterns in data can be exploited in other domains as well.

3 SCUBATEXT PROTOTYPE

We built a prototype to illustrate the idea of integrating textual and graphical presentations in the domain of scuba diving. The architecture of our prototype is a four stage pipeline as shown in Figure 3. The third stage in our architecture performs text generation. This stage will be eventually expanded into the well known three stage NLG pipeline (Rieter and Dale, 2000). With the expanded third stage we will have a six stage pipeline for producing bi-modal presentations of quantitative information.

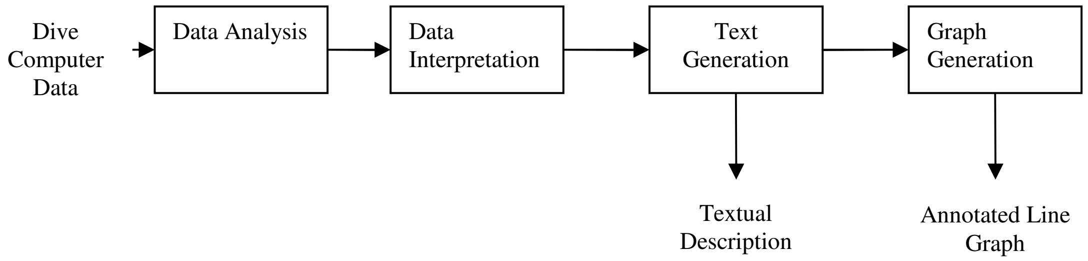


Figure 3: Four stage pipeline architecture of ScubaText Prototype.

Data Analysis - This stage is responsible for determining the structure of the dive (global model) and to detect observable patterns (local patterns) such as ‘stop’ and ‘saw tooth’ that have direct bearing on safety. In other words, this stage performs information extraction from raw dive profile data in terms of local patterns and global models that potentially can be communicated to the user. This stage is largely domain independent except for using domain specific limits for controlling the data analysis algorithms. Most data analysis algorithms are designed to use domain specific data to control their runtime behaviour. The patterns computed in this stage carry all the information required to drive a graphing package to locate them on the line graph and to draw the required annotations. For example, the horizontal line at 40m depth in Figure 1 marks the depth which is 85% of maximum depth of the dive.

Data Interpretation – In this stage, we compute what each of the patterns (or other structural elements) detected in the previous stage ‘mean’ in terms of safety. Here we use a reference dive profile against which we compare the actual dive profile. For the purpose of generating the reference for each dive, we used an algorithm similar to the ones used by dive computers. Dive computers use algorithms known as decompression models developed in the diving medicine community. Decompression models generate recommended safe bottom times and recommended safe ascent schedules given a depth-time profile as input. Using these recommendations, we then compute deviations in the actual dive. Finally we rate a dive based on the deviations: the larger the deviation from the recommended behavior the smaller the rating. In other words, the second stage is where the patterns/models are interpreted using the domain knowledge and also in the context of user tasks. For this purpose analysing dive data for safety related patterns is assumed to be the main user task. Therefore only those patterns/models that are known to be related to dive safety are considered here. The knowledge used in this stage is acquired from several sources. We conducted brief interviews of medical staff working with patients suffering from dive related illnesses to find out the different kinds of patterns linked to DCI. Initially we also carried out some knowledge acquisition (KA) studies with dive instructors at a local diving school. However, these KA studies with medical staff and dive instructors did not produce knowledge sufficient to drive the system development. While building the prototype we relied on dive computer manuals and other electronic sources on the Internet such as news groups to acquire detailed level knowledge. Because we acquired knowledge from several sources we needed to consolidate the knowledge under the guidance of domain experts (Reiter et al., 2003a). However we failed to carry out this knowledge consolidation during the prototype development and as a result we do not claim that our system generated texts can be useful in the real world.

Text Generation – In this stage, we use a classification scheme based on the ratings decided in the above stage to decide the messages to be communicated and also to decide their organization. Based on the overall rating of a dive, we initially generate an overview message about the dive. For example the first sentence in figure 2 ‘Risky dive with some minor problems’ presents an overview message. We then select messages describing those deviations representative of that class and order them in the order of the individual ratings of the deviations. The selected messages are realized using templates. In our prototype, the templates are implemented as procedures that join phrasal/clause fragments based on information such as deviations and ratings computed in the earlier stage.

Graph Generation – In this stage, we determine the graphical support needed to communicate the messages selected in the above stage and pass this information to the graphics package which plots the graph with the required annotations.

4 EVALUATION

In our current work we make two hypotheses:

> Hypothesis 1: Divers find bi-modal (text+graph) presentations more useful to judge the safety of a dive than uni-modal presentations.

> Hypothesis 2: The annotations marked on the graph are useful to interpret the textual description.

We performed a user study to test the above hypotheses. We used sample dive data packaged with a well known model of dive computers in this study. We sent out a questionnaire to divers affiliated to Aberdeen Water Sports, British Sub Aqua Club and Aberdeen University Diver Club and received twenty (20) completed questionnaires. In the questionnaire we showed four different presentations of a sample dive profile data set whose contextual information such as location of the dive, water temperature etc. are provided common to all four presentations:

Presentation A: a line graph showing the dive profile data
Presentation B: a textual description of the analysis of the dive profile data
Presentation C: a combination of the line graph and the textual description of the dive profile data
Presentation D: a combination of the annotated line graph and the textual description cross referencing the line graph

Presentation A is identical to the presentation that the dive computer software allows users to print. Presentation D combines the annotated line graph shown in Figure 1 and the textual description shown in Figure 2. Presentation B contains text similar to the text shown in Figure 2 but with the cross referencing phrases removed. Presentation C combines the line graph from presentation A and the text from presentation B.

In the first question in our questionnaire, subjects (divers) were asked to order the four presentations in the decreasing order of usefulness (most useful first and the least useful last) for judging the safety of a dive. In the second question, they were asked to rank their level of agreement to the following statement on a five point Likert scale (Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree):

> "The annotations on the graph shown in '**Presentation D**' are **useful** for interpreting the textual description."

To test hypothesis 1, we need to show that users find Presentations C and D more useful than Presentations A and B for judging the safety of a dive. For this we collected the ordered list of presentations the twenty participants wrote in response to the first question. Based on the ordered list received from each participant, we ranked presentations with rank values from 4 to 1. The presentation placed first in the ordered list received a rank of 4, the presentation placed next to it received a rank of 3, and the one next to it received 2 and finally the last presentation in the list received a rank of 1. We summed up ranks for A and B to create the rank for uni-modal presentations. Similarly we summed up ranks for C and D to create the rank for bi-modal (text+graph) presentations. We ran the Wilcoxon Matched-Pairs Signed-Ranks Test on these two summed ranks. The test showed (with p <= 1.526e-05) that divers find bi-modal (text+graph) presentations more useful than uni-modal ones. Using this data and using the same statistical test we have tested if divers consider Presentation D more useful than any of the other presentations. The statistical test once again showed that divers indeed consider Presentation D as the most useful one of the four presentations for judging the safety of a dive.

To test hypothesis 2, we need to show that users find the annotations marked in Figure 1 useful for interpreting the text shown in Figure 2 which form the two parts of Presentation D in the questionnaire. For this we collected the Likert point ticked by the participants in response to the second question in the questionnaire. We once again ran the above statistical test on this data set in comparison to the data set obtained with Likert point values expected if all the participants disagreed with the statement in the second question. The test showed (with p <= 5.341e-05) that divers agree that annotations on the graph are useful for interpreting the textual description. In the free text comments at the end of the questionnaire many participants stated that they like the idea of annotations complementing the text and recommended many more annotations that might be useful. Moreover, the participants who disagreed with the statement in question 2, noted in their free text comments that they may change their mind if more useful annotations are marked on the graph. The recommended annotations varied among participants indicating that although users agree with the overall framework used in Presentation D, they need individually tailored implementations of this framework. In our current prototype, we ignored user modeling completely apart from assuming a user task of analyzing dive data for determining the safety of a dive. This single assumption about the user task is used in content selection and organization of text. But there should be more detailed user models to drive the text generation and the graph annotations.

In our questionnaire we have also asked if the participants find the textual description shown in Presentation C (text without cross references to the line graph) appropriate for the line graph shown in the same presentation. Most participants disagreed indicating that our text does not describe the dive data accurately. In the free text comments participants wrote at the end of the questionnaire they indicated that the text is inappropriate because the text is judgmental in its tone. For example, the text says ‘a risky dive’ rather than ‘a potentially risky dive’. In other words, the appropriateness of a text describing a sample data set depends on its emotional content rather than on its factual content. A number of earlier NLG projects such as STOP (Reiter et al., 2003b) and SKILLSUM (Reiter et al., 2005b) faced similar challenges. Participants in our study had concerns about the factual content as well. For example, saw-tooth patterns in the bottom zone need not always be due to ‘poor buoyancy control’ but could be the consequence of the uneven terrain on the sea-bed. In other words, the relationship between a pattern on the dive profile and its interpretation is of type one to many. One of the participants revised the text produced by the system to make it more acceptable both factually and emotionally which is shown in Figure 4. As can be seen the revisions suggested by this participant soften the tone of the text. The saw tooth patterns are not linked to poor buoyancy. Instead they have been linked to a possible occurrence of DCI.

Potentially risky dive with some minor problems. The bottom time of 12.0min exceeds no-stop limit by 4.0min requiring mandatory decompression stops. The ascent was at a constant rate within the recommended rate. The saw tooth patterns marked ‘A’ on the depth-time profile should be avoided if possible as this increases the chance of developing DCI even within the recommended decompression limits. The re-descent from 5m to 10m in the later stages of the dive should also be avoided for the same reason as saw-tooth profiles.

Figure 4: Suggested textual summary of safety related information for the example dive shown in Fig. 1.

5 FUTURE WORK

Our user study showed several problem areas in our current work. The most important one is the absence of user models. Because the current user study showed that bi-modal (text+graph) presentations are useful to users, in future we wish to explore ways of tailoring bi-modal (text+graph) presentations to different users. Another important problem area from the application point of view is to regulate the emotional content of the generated text. Because emotional sensitivity varies from user to user, controlling emotional content too should depend on user models. In our current user study we could not explore the detailed uses of individual annotations. For example, do the markings on the graph showing the ‘bottom zone’ help the user in determining the actual meaning of that phrase? Comprehension questions testing semantic groundings should be included in future questionnaires. Shahar (1997) developed a knowledge based temporal abstraction (KBTA) technique that uses domain knowledge to interpret time series data. Because we aim to produce interpretative summaries of dive computer data, we need to adapt the techniques developed in KBTA. Particularly KBTA makes interpretations that are context dependent. Our evaluation study showed that interpretation of patterns in our application too is context dependent. In the future we want to explicitly represent dive context and reason with it for a more acceptable data interpretation.

6 CONCLUSION

In this paper, we present a bi-modal (text+graph) scheme for presenting summaries of quantitative information. The graphical part of the presentation communicates the un-interpreted information and also helps the user in interpreting portions of text that refer to structural elements in the data. The textual part of the presentation communicates interpreted information. Thus both the modes are linked to each other; the graphical part depends upon the textual description for its proper interpretation and certain portions of the textual part depend on the graphical part for their interpretation. An evaluation study showed that users find our bi-modal (text+graph) presentation scheme more useful than the uni-modal presentations. Also the evaluation study showed that using annotations for semantically grounding phrases referring to patterns in the input data actually helps users interpret texts. We believe that the idea of semantic grounding explored in this work will be relevant in other domains that address the problem of communicating information related to non-standard patterns to the user.

ACKNOWLEDGEMENTS

Many thanks to all the divers/instructors who participated in our user study for taking time to write a lot of comments and suggestions. Many thanks to Dr John Ross, Occupational Medicine, Aberdeen University who originally suggested the SCUBA domain to apply techniques developed in the SumTime project. This work is supported by The Nuffield Foundation under their grant scheme for newly appointed science lecturers (NAL/00889/G).

REFERENCES

André, E. (2000). The generation of multimedia presentations. In: *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text* , R. Dale, H. Moisl, and H. Somers (Eds.), 305-327, Marcel Dekker Inc., 2000.

Baldonado, M.Q.W., Woodruff, A. and Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. In *Proceedings of ACM Advanced Visual Interfaces '00.*

Boyd, S. (1997). Detecting and describing patterns in time-varying data using wavelets. In: *Advances in Intelligent Data Analysis: Reasoning About Data*, X. Lui and P. Cohen (Eds.), Lecture Notes in Computer Science 1280, Springer Verlag.

Dasgupta, D. and Forrest, S. (1996). Novelty detection in time series data using ideas from immunology. *Proceedings of the 5th International Conference on Intelligent Systems*, Reno.

Fasciano, M. and Lapalme, G. (1996). PostGraphe: a system for the generation of statistical graphics and text. *Proceedings of the 8th International Workshop on Natural Language Generation (INLG '96)*, pp 51-60.

Fayyad, U., Grinstein, G.G. and Wierse, A. (2002). *Information Visualization in Data Mining and Knowledge Discovery*. Academic Press.

Hunter, J.R.W., Freer, Y., Ewing, G., Logie, R., McCue, P. and McIntosh, N. (2003). NEONATE: Decision support in the neonatal intensive care unit – A Preliminary Report. *AIME-03: Proceedings of the Ninth European Conference on Artificial Intelligence in Medicine*, Springer Verlag, pp 41-45.

Keogh, E., Chu, S., Hart, D. and Pazzani, M. (2001). An online algorithm for segmenting time series. *Proceedings of IEEE International Conference on Data Mining*, pp 289-296.

Keogh, E., Lonardi, S and Chiu, W. (2002). Finding surprising patterns in a time series database in linear time and space. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada, pp 550-556.

Keogh, E. Lin, J. and Fu, A. (2005). HOT SAX: Efficiently finding the most unusual time series subsequence. *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005)*, pp 226 - 233.

Law, A.S., Freer, Y., Hunter, J.R.W., Logie, R.H., McIntosh, N. and Quinn, J. (2005). A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing* 19:183-194.

Langan-fox, J., Platania-phung, C. and Waycott, J. (2006). Effects of advance organizers, mental models and abilities on task and recall performance using mobile phone network. *Applied Cognitive Psychology*, 20: 1143-1165.

Lin, J., Keogh, E., Patel, P. and Lonardi, S. (2002). Finding motifs in time series. *Proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada.

Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., and Shneiderman, B. (1998). LifeLines: Using visualization to enhance navigation and analysis of patient records. Revised version in 1998 *American Medical Informatic Association Annual Fall Symposium*, pp. 76-80.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Reiter, E., Sripada, S. and Robertson, R. (2003a). Acquiring correct knowledge for Natural Language Generation, *JAIR*, 18:491-516.

Reiter, E., Robertson, R., and Osman, L. (2003b). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence* 144:41-58.

Reiter E., Sripada, S., Hunter, J., Yu, J. and Davy, I. (2005a). Choosing words in computer-generated weather forecasts. *Artificial Intelligence* 167:137-169.

Reiter, E., Williams, S. and Crichton, L. (2005b). Generating feedback reports for adults taking basic skills tests. In: *Applications and Innovations in Intelligent Systems XIII (Proceedings of ES-05)*, A. Macintosh, R. Ellis and T. Allen (Eds), pp 50-63.

Shahar, Y. (1997), Framework for knowledge-based temporal abstraction. *Artificial Intelligence* 90:79-133.

Shneiderman, B. (1996) The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the 1996 IEEE Conference on Visual Languages,* Boulder, CO, pp 336-343.

Sripada, S. Reiter, E. and Davy, I. (2003a). SumTime-Mousam: configurable marine weather forecast generator. *Expert Update* 6(3):4-10.

Shneiderman, B. and Plaisant, C. (2005). *Designing User Interfaces*. Addison-Wesley.

Sripada, S. Reiter, E. Hunter, J. and Yu, J. (2003b). Summarising neonatal time series data. *Proceedings of the 2003 Conference of the European Chapter of the Association for Computation Linguistics*, Companion Volume, pp 167-170.

Theune, M. (2001). ANGELICA: choice of output modality in an embodied agent. *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD-2001)*, Verona, Italy, pp 89-94.

Tufte, E. (2001). *The Visual Display of Quantitative Information*. Second Edition, Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Weber, M., Alexa, M. and Müller W. (2001). Visualizing time-series on spirals. *Proceedings of the IEEE Symposium on Information Visualization 2001* (INFOVIS'01).

Yu, J., Reiter, E., Hunter, J. and Mellish, C. (2006). Choosing the content of textual summaries of large time-series data sets. To appear in *Natural Language Engineering*.

List of authors