SEMDIAL 2015
goDIAL

Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue

Christine Howes and Staffan Larsson (eds.)

Gothenburg, 24–26 August 2015

UNIVERSITY OF
GOTHENBURG

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

http://www.illc.uva.nl/semdial/

goDIAL Website

http://www.flov.gu.se/godial

goDIAL Sponsors

CLT

Talkamatic
FREE DIALOGUE

Swedish Research Council
Vetenskapsrådet

goDIAL Endorsements

SIGdial
Special Interest Group on Discourse and Dialogue

Preface

goDIAL brings the SemDial Workshop on the Semantics and Pragmatics of Dialogue back to the University of Gothenburg, where the fourth meeting - GÖTALOG - took place in 2000. goDIAL, and the SemDial workshop as a whole, is unique in offering a cross section of dialogue research including experimental studies, corpus studies, and formal models of the semantics and pragmatics of dialogue.

We received a total of 40 full paper submissions, 17 of which were accepted after a peer-review process, during which each submission was reviewed by a panel of three experts. We are extremely grateful to the Programme Committee members for their very detailed and helpful reviews. The poster session hosts 8 of the remaining submissions, together with 21 additional submissions that came in response to a call for late-breaking posters and demos. All accepted full papers and poster abstracts are included in this volume.

The goDIAL programme features three keynote presentations by Ellen Bard, Elisabet Engdahl and Marilyn Walker. We thank them for participating in SemDial and are honoured to have them at the workshop. Abstracts of their contributions are also included in this volume.

goDIAL has received generous financial support from the Swedish Research Council (`http.//www.vr.se`), Talkamatic (`http://www.talkamatic.se`), and the Centre for Language Technology at Gothenburg University (`http://clt.gu.se`). We are very grateful for this sponsorship. We have also been given endorsements by the ACL Special Interest Groups: SIGdial and SIGSEM.

This year, we are very happy and proud to co-locate goDIAL with the inauguration workshop of CLASP (`http://flov.gu.se/clasp/inauguration-workshop`), the Centre for Linguistic Theory and Studies of Probability.

Last but not least we would like to thank our local organisers Simon Dobnik and Ellen Breitholtz for their tireless work, and everyone else who helped with all aspects of the organisation, including our student helpers.

Christine Howes and Staffan Larsson

Gothenburg

August 2015

Programme Committee

Nicholas Asher	IRIT, CNRS
Claire Beyssade	CNRS/Institut Jean Nicod
Ellen Breitholtz	University of Gothenburg
Sarah Brown-Schmidt	University of Illinois
Eve Clark	Stanford University
Liz Coppock	University of Gothenburg
Chris Cummins	University of Edinburgh
Valeria De Paiva	University of Birmingham
Paul Dekker	ILLC, University of Amsterdam
David DeVault	ICT USC
Simon Dobnik	University of Gothenburg
Arash Eshghi	Heriot Watt University
Raquel Fernandez	University of Amsterdam
Kallirroi Georgila	University of Southern California
Jonathan Ginzburg	Universite Paris-Diderot, Paris 7
Eleni Gregoromichelaki	King’s College London
Pat Healey	Queen Mary University of London
Anna Hjalmarsson	KTH
Judith Holler	Max Planck
Julian Hough	Bielefeld University
Chris Howes	University of Gothenburg
Amy Isard	University of Edinburgh
Kristiina Jokinen	Univerity of Helsinki
Ruth Kempson	King’s College London
Alexander Koller	University of Potsdam
Staffan Larsson	University of Gothenburg
Alex Lascarides	University of Edinburgh
Pierre Lison	University of Oslo
Peter Ljunglöf	Chalmers Institute of Technology
Colin Matheson	University of Edinburgh
Gregory Mills	University of Groningen
Chris Potts	Stanford University
Matthew Purver	Queen Mary University of London
Hannes Reiser	Bielefeld University
David Schlangen	Bielefeld University
Gabriel Skantze	KTH
Matthew Stone	Rutgers
David Traum	ICT USC

Table of Contents

Invited Talks

How weird is that? Predictability and cognitive difficulty in dialogue 2
Ellen Bard

How to connect an utterance: Strategies for cohesive dialogues in Scandinavian 3
Elisabet Engdahl

Semantics and sarcasm in online dialogue 4
Marilyn Walker

Oral Presentations

Taking a stance: A corpus study of reported speech 6
Shauna Concannon, Pat Healey and Matthew Purver

Shifting opinions: Experiments on agreement and disagreement in dialogue 15
Shauna Concannon, Pat Healey and Matthew Purver

Changing perspective: Local alignment of reference frames in dialogue 24
Simon Dobnik, Christine Howes and John Kelleher

Learning non-cooperative dialogue policies to beat opponent models: “The good, the bad and the ugly” 33
Ioannis Efstathiou and Oliver Lemon

Exploring age-related conversational interaction 42
Jeroen Geertzen

Engagement driven topic selection for an information-giving agent 48
Nadine Glas, Ken Prepin and Catherine Pelachaud

Building and applying perceptually-grounded representations of multimodal scene descriptions . . 58
Ting Han, Casey Kennington and David Schlangen

User information extraction for personalized dialogue systems 67
Toru Hirano, Nozomi Kobayashi, Ryuichiro Higashinaka, Toshiro Makino and Yoshihiro Matsuo

How far can we deviate from the performative formula? 77
Lisa Hofmann

Timing and grounding in motor skill coaching interaction: Consequences for the information state 86
Julian Hough, Iwan de Kok, David Schlangen and Stefan Kopp

Defining the right frontier in multi-party dialogue 95
Julie Hunter, Nicholas Asher, Eric Kow, Jérémy Perret and Stergos Afantenos

Learning trade negotiation policies in strategic conversation 104
Simon Keizer, Heriberto Cuayáhuitl and Oliver Lemon

Reducing the cost of dialogue system training and evaluation with online, crowd-sourced dialogue data collection 113

Ramesh Manuvinakurike, Maike Paetzel and David Devault

When hands talk to mouth: Gesture and speech as autonomous communicating processes 122
Hannes Rieser

Interpreting English pitch contours in context 131
Julian Schlöder and Alex Lascarides

"Hand me the yellow stapler" or "hand me the yellow dress": Colour overspecification depends on object category 140
Sammie Tarenskeen, Mirjam Broersma and Bart Geurts

Editing Phrases 149
Ye Tian, Claire Beyssade, Yannick Mathieu and Jonathan Ginzburg

Poster Presentations

The significance of silence: Long gaps attenuate the preference for 'yes' responses in conversation 158
Sara Bögels, Kobin Kendrick and Stephen Levinson

Within reason: Categorising enthymematic reasoning in the balloon task 160
Ellen Breitholtz and Christine Howes

Context influence vs efficiency in establishing conventions: Communities do it better 162
Lucia Castillo, Holly Branigan and Kenny Smith

Are response particles not well understood? Yes/No, they arent! An experimental study on German ja and nein 164
Berry Claus, Marlijn Meijer, Sophie Repp and Manfred Krifka

KILLE: Learning objects and spatial relations with kinect 166
Erik de Graaf and Simon Dobnik

Demonstrating the dialogue system of the intelligent coaching space 168
Iwan de Kok, Julian Hough, Felix Hülsmann, Thomas Waltema, Mario Botsch, David Schlangen and Stefan Kopp

Non-sentential utterances in dialogue: Experiments in classification and interpretation 170
Paolo Dragone and Pierre Lison

DS-TTR: An incremental, semantic, contextual parser for dialogue 172
Arash Eshghi

What could interaction mean in natural language and how could it be useful? 174
Christophe Fouqueré and Myriam Quatrini

Towards the automatic extraction of corrective feedback in child-adult dialogue 176
Sarah Hiller and Raquel Fernandez

Spot the difference - A dialog system to explore turn-taking in an interactive setting 178
Anna Hjalmarsson and Margaret Zellers

Eye blinking as addressee feedback in face-to-face conversation 180
Paul Hömke, Judith Holler and Stephen C. Levinson

A reusable interaction management module: Use case for empathic robotic tutoring 182
Srinivasan Janarthanam, Helen Hastie, Amol Deshmukh, Ruth Aylett and Mary Ellen Foster

Concern-alignment analysis of consultation dialogues 184
Yasuhiro Katagiri and Katsuya Takanashi

The interactive building of names 186
Ruth Kempson, Stergios Chatzikyriakidis and Ronnie Cann

Playing a real-world reference game using the words-as-classifiers model of reference resolution . 188
Casey Kennington, Soledad Lopez Gambino and David Schlangen

The state of the art in dealing with user answers 190
Staffan Larsson

Polish event-linking devices of przed 'before' cluster in conversational data Implications for contrastive analysis 192
Barbara Lewandowska-Tomaszczyk

Developing spoken dialogue systems with the OpenDial toolkit 194
Pierre Lison and Casey Kennington

CCG for discourse 196
Sumiyo Nishiguchi

Experimenting with grounding strategies in dialogue 198
Volha Petukhova, Harry Bunt, Andrei Malchanau and Ramkumar Aruchamy

Self awareness for better common ground 200
Ondřej Plátek and Filip Jurčíček

Adaptive dialogue management in the KRISTINA project for multicultural health care applications 202
Louisa Pragst, Stefan Ultes, Matthias Kraus and Wolfgang Minker

Toward a scary comparative corpus: The Werewolf spoken corpus 204
Laurent Prévot, Yao Yao, Arnaud Gingold, Bernard Bel and Kam Yiu Joe Chan

Towards a layered framework for embodied language processing in situated human-robot interaction 206
Matthias A. Priesters, Malte Schilling and Stefan Kopp

How questions and answers cohere 208
Mandy Simons

Modeling referential coordination as a particle swarm optimization task 210
Hodgdon Stevens and Hannah Rohde

Distribution of non-sentential utterances across BNC genres: A preliminary report 212
Kwong-Cheong Wong and Jonathan Ginzburg

Interactive learning through dialogue for multimodal language grounding 214
Yanchao Yu, Oliver Lemon and Arash Eshghi

Invited Talks

How weird is that? Predictability and cognitive difficulty in dialogue

Ellen Bard
University of Edinburgh
`ellen@ling.ed.ac.uk`

Many psycholinguistic processes appear to be sensitive to the probabilities of the available choices with extra processing, production, or learning where an option's probability is low. This talk will discuss evidence for a similar principle within the cognitive difficulty of dialogue. Using designed but unscripted dialogue corpora, I will show 1) that when dialogue structure demands sequences of low predictability, difficulty rises; 2) that priming in dialogue, a force for increasing the probability of matching actions in adjacency pairs, is very broadly based; 3) but that its effects may be limited, in particular, by differences between interlocutors' tasks. Finally, I will spend some time showing how a proprietary dialogue protocol is designed to control predictability in uncertain and dangerous circumstances.

How to connect an utterance: Strategies for cohesive dialogues in Scandinavian

Elisabet Engdahl
University of Gothenburg
`elisabet.engdahl@svenska.gu.se`

Many syntacticians consider preposing to an utterance initial position be a marked option in the grammar which has to be licensed by some discourse function, as in the case of questions (1) or so-called topicalization (2).

(1) What did she say?

(2) That/*it I don't like. (I'd rather have some ...)

In English, topicalized constituents are normally stressed and invoke a notion of contrast; an unstressed personal pronoun is not felicitous. In the mainland Scandinavian languages, Danish, Norwegian and Swedish, preposing of unstressed pronouns is quite common as a way to connect an utterance to the preceding context, as illustrated by the Swedish example in (3).

(3) *A: Var är cykeln?*
where is bike-DEF
'Where is the bike?'

B: Den ställde jag i garaget.
it put I in garage-DEF.
'I put it in the garage.

In order to find out when this type of preposing is used in dialogue, Filippa Lindahl and I carried out a search in the Nordic Dialect Corpus, a 2.5 million word corpus of spontaneous conversations Johannessen et al. (2009). In my talk I will present some common strategies that we found and discuss their relevance for both syntactic theory and spoken dialogue systems.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., and Vangsnes, Ø. A. (2009). The nordic dialect corpus-an advanced research tool. In Jokinen, K. and Bick, E., editors, *Proceedings of the 17th Nordic conference of computational linguistics NODALIDA 2009. NEALT proceedings series*, volume 4, pages 73–80.

Semantics and Sarcasm in Online Dialogue

Marilyn Walker
University of California Santa Cruz
`maw@soe.ucsc.edu`

Online forums provide a fascinating source of data for research on the structure of dialogue. Unlike traditional media corpora, online conversation is highly social and subjective and its interpretation and analysis are strongly dependent on context. Phenomena such as sarcasm and rhetorical questions abound. In this talk I will first describe the IAC corpus that we have made publicly available. I will then discuss our research on several tasks related to dialogue structure and the meaning of utterances, such as recognizing sarcasm, distinguishing agreement from disagreement, identifying the linguistic properties of factual vs. emotional arguments, and mining the aspects of arguments on different topics.

Oral Presentations

Taking a Stance: a Corpus Study of Reported Speech

Shauna Concannon **Patrick G. T. Healey** **Matthew Purver**
Queen Mary University of London
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
{s.concannon; p.healey; m.purver}@qmul.ac.uk

Abstract

People tend to avoid exposed disagreement in conversation. This is normally attributed to politeness strategies that mitigate the potential face-threat created by direct disagreement with a conversational partner. In reported speech the pressure for mitigation of negative responses is removed, leading to the prediction that reported speech should contain more exposed disagreement. However, concerns about self-presentation may lead people to present their prior behaviour in such a way that demonstrates their understanding that disagreement is a sensitive matter; thus, differences in self-reported and other-reported disagreement would be anticipated. Finally, we predict that reported speech is used to highlight substantive differences in stance, and contains more explicit markers of stance to highlight newsworthiness. To test these ideas we compare the distribution of markers of agreement, disagreement and stance in four samples of conversation from the BNC: direct speech, self-reported speech (I said), other-reported speech (he / she said) and local dialogue context. Contrary to the prediction the results show that both direct and indirect markers of agreement and disagreement are more common in direct speech than reported speech. However, markers of contrast and emphasis including negations, swearwords and contrastive conjuncts are both more common in reported speech than direct speech and in self-reported speech than other-reported speech.

1 Introduction

In spoken dialogue people sometimes talk about things that were said in other conversations. These instances of reported speech are typically marked by a pronoun (e.g., 'he', 'she', 'I') and an embedding verb (e.g., 'said', 'went', 'goes') followed by a rendition of the previous utterance, as demonstrated by the following examples, taken from the British National Corpus:[1]

> *I said, I'm not assassinating your character now but you're being very intimidating in the way that your talking to people.*[2]
>
> *So she said, well you can't do that.* [3]

Example 1

Detailed studies of the form and function of reported speech show that they are not simple verbatim reproductions of something said previously (Clark and Gerrig, 1990; Clift, 2006; Clift, 2007; Holt, 2000; Holt, 2007). Rather, they involve the selective representation of people's own and others' conversational conduct. This allows conversational participants to use them, amongst other things, as evidence or justification for particular accounts of events, to relay complaints and disputes and to claim *epistemic priority* or privileged rights, knowledge or expertise about a topic under discussion (Holt, 2000; Clift, 2006; Haakana, 2007; Vincent and Perrin, 1999). The non-narrative functions of reported speech have been closely associated with the expression of a *point of view* and argumentation, providing justification,

---

[1] Data cited herein have been extracted from the British National Corpus, distributed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

[2] Theatre public meeting, September 1991, BNC-D91

[3] At home, March 1992, BNC-KCN1

support or authority for a particular stance (Vincent and Perrin, 1999; Couper-Kuhlen, 2007). It has been noted that reported speech is often more blunt or forthright in character, and constructed in such a way that the reported speech, and the action performed by it, is easily recognisable (Clift, 2007).

The difference between what is said and what is reported as said thus provides a potentially useful analytic window on the specific ways people use language to produce these different pragmatic effects. Here we focus in particular on what this contrast can tell us about the way people formulate and report on their agreements and disagreements with others.

Direct challenges and disagreement in conversation are socially problematic. As we discuss below, exposed disagreement is generally avoided (Pomerantz, 1977) because it is potentially face threatening (Brown and Levinson, 1987). If people are reluctant to expose disagreements directly then reported speech provides a potentially useful context in which prior disagreements could be presented more explicitly; the original addressee is absent which reduces concerns about politeness and the likelihood of a challenge to the speaker's version of events.

What can reported speech tell us about the differences between how people enact disagreement and how they represent their disagreements in conversation? Which elements are preserved in the representation of (dis)agreement and which are not? To address these questions we test whether there are systematic differences in the manifestation of agreement and disagreement in direct speech and reported speech in a large corpus of everyday conversations (Burnard, 1995). In particular, we look at the distribution of markers of (dis)agreement, updates, contrast and emphasis. We compare how people use these in direct speech, in reports of their own speech and the speech of others.

The paper proceeds by briefly setting out qualitative, conversation analytic (CA), research on how disagreements are typically managed in direct conversation. We then consider the different markers of (dis)agreement, contrast and stance that can be used to inform a quantitative analysis. This enables a comparison of the ways people both enact and report on their agreements and disagreements. We compare, in particular, a) direct speech with reported speech b) self-reported speech ('I said') with other-reported speech ('He said', 'She said') and, in order to check effects of conversational context, c) self-reported speech with direct speech by the same speaker in their talk immediately preceding the reported speech.

The results show, contrary to our predictions, that explicit agreement and disagreement are more common in direct speech than reported speech. Nonetheless, markers of contrast and emphasis including negations, swearwords and contrastive conjuncts are both more common in reported speech than direct speech and more common in self-reported speech than other-reported speech. We propose that people use reported speech primarily to present the substance of their differences with a prior addressee rather than to re-present how those differences were played out.

1.1 Avoiding Disagreements

Making and responding to assessments and other assertions is a common feature of conversation. Conversation analysts have shown that when people produce initial assessments of situations or events, positive responses are made more quickly and clearly than negative or unaligned responses (Sacks, 1987; Pomerantz, 1977). Negative or *dispreferred* responses are normally produced more slowly, are often prefaced with some form of agreement ('Oh yes... but') and the negative assessment itself is often delayed by several turns and produced with some sort of mitigating account (Pomerantz, 1977).

When responding to an initial assessment, an agreement may be signalled by repeating back the original assessment, but whether this is an exact repeat or a modified repeat can signal whether it is a strong agreement or weaker variation, acting to modify or downgrade an assessment or perhaps even disagree. In the following example, taken from Pomerantz (1977), pauses and delays, such as the '(hhhhh)', may suggest the speaker is taking some time to formulate their disagreement, or decide upon the most tactful way to deliver it:

A: cause those things take working at,
(2.0)
B: (hhhhh) well, they do, but
A: They aren't accidents,
B: No, they take working at, But on the other hand, some people are born with uhm (1.0)

B: well a sense of humor, I think it's something you are born with Bea.

A: Yes. Or it's c- I have the- eh yes, I think a lotta people are, but then I think it can be developed too.

Example 2

In addition to the hesitation, speaker B also uses the discourse marker *well*, often used to highlight that a disagreement is forthcoming. Furthermore, speaker A performs an initial agreement by repeating back *they take working at*, before delivering a contrasting point of view, namely that certain traits are innate. In response speaker A also offers an appeasing agreement, before reverting back to their previous, contrary stance, *I think it can be developed too*. This small extract highlights many of the devices, such as hesitation, negation, and discourse markers, that are employed when managing disagreement in dialogue.

The CA observations highlight the ways that people normally avoid exposing disagreements directly (unless of course they intend to be abrupt or confrontational). Consequently explicit markers of disagreement should tend to be rare in conversation and much less common than explicit markers of agreement. How would we expect these phenomena to play out in reported speech?

1.2 Hypotheses

We distinguish three general hypotheses for reported speech:

1. Politeness: The general politeness hypothesis is that people avoid the face-threat involved in direct disagreement with an addressee. Unless a current addressee is aligned in some way with the person(s) whose speech is being reported then the pressure for mitigation of negative responses is removed.[4] The general *politeness hypothesis* thus predicts that reported speech should tend to contain more exposed disagreement than direct speech.

2. Self-Presentation: Even where people are not disagreeing directly with their current addressee they might still wish to demonstrate that they understand that disagreement is a sensitive matter e.g., to avoid the inference that they are rude or combative. If people are sensitive to this then, all things being equal, they should not produce any more explicit disagreements in reported speech than they do in direct speech. Moreover, concerns about self-presentation should by definition affect 'self 'more strongly than 'other' therefore we would expect fewer explicit markers of disagreement in self-reported than other-reported speech.

3. Contrastive Stance: A third general hypothesis is that people's primary concern when reporting on a prior conversation is to highlight the substantive differences between their own stance and that of others. The intuition here is that like ordinary utterances reported speech should ideally be newsworthy in some way (Goodwin, 1979); either to the current addressee as a means of highlighting a significant stance previously taken by the speaker, or to convey the newsworthiness of the reported speech to the people actually in the prior conversation. This leads to the prediction that reported speech should contain more explicit markers of stance or emphasis than direct speech; for example, by using turn-initial discourse markers such as 'well' or negations (Scott, 2002) as illustrated above in Example 1.

In order to make quantitative tests of these predictions we now consider in more detail some potential indices of the different ways people can position direct and reported speech. In particular, discourse markers of (dis)agreement, stance, emphasis and contrast.

1.3 Markers of Agreement and Disagreement

The simplest case for analysis is where people explicitly position their turns as agreement or disagreement. This can be done with phrases such as 'You're wrong', 'I disagree', 'I don't agree' and 'You're right' or 'I agree'. Unfortunately, for the reasons outlined above these exposed forms, especially those associated with disagreement, are likely to be rare.

A second set of more indirect indicators are provided by cue words or discourse markers that are associated with agreement and disagreement but don't explicitly formulate a turn as such. Walker et al. (2012) analysed large datasets of forum posts to identify cue words marking features such as agreement, disagreement and sarcasm. Samples were manually annotated for levels of disagreement and agreement. In order of decreasing consensus amongst annotators the markers of dis-

---

[4] Of course, it is possible that the current addressee might also take issue with the opinion or stance identified in the reported speech but this would become an issue for their *subsequent* response to the report not the format of the report itself.

agreement were: ‘really’ (67% read a response beginning with this marker as prefacing a disagreement with a prior post), ‘no’ (66%), ‘actually’ (60%), ‘but’ (58%), ‘so’ (58%), and ‘you mean’ (57%).

These markers do not, of course, encompass all ways of doing disagreement. About 50% of respondents interpreted unmarked posts as disagreeing, highlighting the way disagreement is often enacted by more indirect means. Walker et al. (2012) also identified markers of agreement: ‘yes’ (73% read a response beginning with this marker as prefacing an agreement), ‘I know’ (64%), ‘I believe’ (62%), ‘I think’ (61%), and ‘just’ (57%).

One limitation of these indirect markers is that they are drawn from analysis of online discussion forums which are less dialogical than face-to-face interaction and where people may also tend to actively seek out disputes. It is also worth noting that, for example, the frequency of turn-initial ‘yes’ is not an unambiguous indicator of agreement; disagreement is often preceded by techniques including agrees (e.g. ‘yes, but...’), delays and prefaces, such as, ‘well’ and ‘hmm’ (Sacks and Jefferson, 1995; Pomerantz, 1977; Kotthoff, 1993). Clift (2006) observes that ‘well’ can act as a buffer. Nonetheless, we assume that the relative distribution of these markers across different samples is indicative of the overall patterns of agreement and disagreement within them.

1.4 Update Markers

In addition to marking the fact of agreement and disagreement there are more subtle pragmatic markers that can signal an individual’s knowledge state or stance with respect to the current conversational context. Here we use ‘well’ and ‘oh’ , which we gloss as *update markers* both of which are associated with signalling some form of contrast or sequential discontinuity in dialogue.

A turn-initial ‘well’ typically (but not exclusively) indicates that what follows will be in some way unexpected, unwelcome, discontinuous or contrary to a prior statement (Pomerantz, 1984; Schegloff and Lerner, 2009; Schiffrin, 1988; Heritage and Clayman, 2010). As such it can signal a forthcoming utterance, that is contrasting, unexpected or perhaps unwanted in substance, and which will lead to an update of the knowledge status.

A turn-initial *oh*, by contrast, typically (but not exclusively) acts as a reactive *change-of- state* token that indexes a responsive shift to a prior utterance through an update in the speaker’s knowledge or awareness (Heritage, 1984; Heritage, 1998). Schiffrin (1988) observes that *oh* often marks a shift in speaker orientation or stance, indicating a speaker’s realisation that the hearer is not similarly aligned or oriented towards a proposition and may signal a potentially argumentative stance.

1.5 Contrast, Emphasis and Expletives

Finally, in order to index the way in which the content of a turn is formulated or positioned with respect to another turn, we track negations (‘not’ and ‘n’t’) and mid-turn contrastive conjuncts (‘but’ and ‘though’) as markers of contrast. The role of negation as a key phenomenon in relation to opinion and disagreement has been noted in the literature (Scott, 2002; Benamara et al., 2012) and is of particular interest here because of its use for the denial or rejection of statements; consequently, its role in rejection and disagreement, together with its inherent connection to the expression of alternatives or contrast, led to the inclusion of negation for our analysis. Adverbial emphasisers, such as ‘really’, ‘surely’ and ‘clearly’, are included as indicators of emphasis (Quirk and Crystal, 1985). The role of adverbial emphasisers as possible indices of disagreement (Scott, 2002) and for the expression of stance (e.g. conveying attitudes towards the content of a sentence), have been highlighted in the literature (Biber and Finegan, 1989; Conrad and Biber, 2000). We also track frequencies of a manually compiled list of common swearwords informed by previous studies and frequency data that surfaced those common to the BNC dataset (‘bastard’, ‘bitch’, ‘bloody’, ‘bollocks’, ‘fuck’, ‘piss off’, ‘shit’ and ‘wanker’) which can be used for the expression of emotions, especially frustration, anger and surprise (Jay and Janschewitz, 2008).

2 Predictions

Building on the three general hypotheses presented above and the discussion of different markers of agreement, disagreement and stance we can summarise eight basic predictions:

1. Politeness: Markers of agreement should always be more common than markers of disagreement in all speech.

2. Politeness: Markers of disagreement should be more common in reported speech than direct speech.

3. Politeness: Expletives should be more common in reported speech than direct speech.

4. Self-Presentation: Markers of disagreement should not be more common in self-reported speech than direct speech.

5. Self-Presentation: Markers of disagreement should be less common in self-reported speech than other-reported speech.

6. Self-Presentation: Expletives should be less common in self-reported speech than other-reported speech.

7. Contrastive Stance: Update markers should be more common in reported speech than in direct speech.

8. Contrastive Stance: Contrast and Emphasis should be more common in reported speech than in direct speech.

3 Method

The corpus analysis used the spoken dialogue component of the British National Corpus (BNC), comprising approximately 10 million words. This sizeable collection of naturally occurring conversations offers scope to explore patterns of reported speech across a large sample. The transcripts include annotations for some key paralinguistic features such as laughing, overlapping speech and significant pauses, although the transcription conventions vary. Our analysis is based on the BNC's *s-units* which are sentence-like divisions of the transcribed utterances. We used SCoRE, a web interface for dialogue corpora, to gather our data from the BNC (Purver, 2001). It can be used to search for any regular expression, and for word or phrase repetitions, including repeats across sentence/turn boundaries.

For each set of markers their frequency in the BNC was gathered and analysed. Reported speech can be introduced in a number of ways, for example, 'I went', 'I says', 'he goes', 'she was like'. We focused on 'pronoun + said + report' as with produced a good sized dataset. Using the ScoRE interface (Purver, 2001) it was possible to extract all instances of 'I said' (5315 turns), 'he said' (3310 turns) and 'she said' (2579 turns), which were then checked by hand to ensure they were consistent samples of reported speech. A further 5315 turns were randomly selected from the spoken dialogue section of the BNC to provide a comparable sample of general direct speech.

In order to control for the possibility that reported speech tends to occur in particular dialogue contexts or with particular audiences (e.g., storytelling to friends) a second sub-sample of 500 turns of direct speech was selected from the same context by identifying the nearest preceding turn to an identified instance of self-reported speech ('I said') by the same speaker, that did not contain an instance of reported speech. This is referred to below as the *Local Context* sample.

The samples were analysed for a number of turn-initial features: agreement and disagreement markers, update markers 'oh' and 'well'. Turn-initial in the reported speech samples constituted what immediately followed I/(s)he said, while in the direct speech sample it was simply the initial words of the turns. Non-turn-initial features were also investigated: adverbial emphasisers (often indicators of stance or opinion markers), 'oh' (change-of-state tokens), negations and swearwords.

4 Results

4.1 Exposed Disagreement

As Table 1 shows, both exposed agreement and disagreement are rare, although exposed agreement is, as expected, more common than disagreement. Only 0.8% of the turns sampled contain strong expressions of disagreement whereas 5.2% contain strong expressions of agreement. Strikingly, over 97% of these instances of exposed agreement/disagreement occur in direct speech. This observation is clearly counter to the initial politeness hypothesis for reported speech and incompatible with the self-presentation hypothesis.

Chi Square analysis of the frequency of strongly exposed agreement and disagreement indicates that their distributions are different in reported and direct speech ($\chi^2_{(1)} = 15.23$, $p<0.01$). [5] There is approximately a 7:1 bias toward overt expression of agreement over disagreement in direct speech compared with approximately 1:1 in re-

---

[5] Throughout we use $p<0.05$ as our criterion level but report computed probabilities to two decimal places for completeness.

Phrase	RS	DS	Total
You're wrong	6	17	23
I disagree	0	15	15
I don't agree	2	46	48
You're right	5	224	229
I agree	5	318	323

Table 1: Instances of Exposed Agreement and Disagreement in the BNC. *RS* = Reported Speech and *DS* = Direct Speech

ported speech. This suggests that although explicit, exposed disagreement is much less common in reported speech there is no particular bias in that context toward overtly positioning a relayed turn as agreement or disagreement.

4.2 Agreement and Disagreement markers

The distribution of turn-initial markers of agreement and disagreement identified by (Walker et al., 2012) for each subsample are shown in Tables 2 and 3.

Marker	DS	(s)he said	I said	Context
Really	3	5	4	1
No	173	128	190	12
Actually	3	5	2	1
But	85	63	51	13
So	108	30	17	19
You mean	0	0	0	0
Total	**372**	**231**	**264**	**46**
Total turns	5315	5889	5315	500
% total turns	7.00	3.92	5.00	9.20

Table 2: Frequency of Disagreement Markers

As Table 2 suggests, the overall frequency of markers of disagreement is higher in direct speech than all reported speech ($\chi^2_{(1)}$ = 48.3, p<0.01) and also higher in the Local Context sample (i.e. preceding direct speech turn by the same speaker) than in the self-reported speech of the same speaker ($\chi^2_{(1)}$ = 16.22, p<0.01). Comparison of self-reported speech with other-reported speech (he/she said) shows markers of disagreement are less common in other-reported speech ($\chi^2_{(1)}$ = 7.22, p=0.01). These patterns are opposite to the predicted pattern for the Politeness and Self-Presentation hypotheses for reported speech.

Marker	DS	(s)he said	I said	Context
Yeah/Yes	647	139	181	26
I know	12	16	22	4
I believe	0	1	1	1
I think	31	22	27	3
I just	4	10	6	2
Total	**694**	**188**	**237**	**36**
Total turns	5315	5889	5315	500
% total turns	13.06	3.19	4.46	7.20

Table 3: Frequency of Agreement Markers

The same pattern is observed for the markers of agreement. They are more common in direct than reported speech ($\chi^2_{(1)}$ = 489, p<0.01) and more common in the Local Context sample from the same speaker than in self-reported speech ($\chi^2_{(1)}$ = 7.63, p=0.01). They are also more common in self-reported speech than other-reported speech ($\chi^2_{(1)}$ = 12.2, p<0.01).

Overall the results show that explicit and implicit markers of agreement and disagreement are more common in direct speech than reported speech and more common in self-reported than other-reported speech.

4.3 Turn-Initial Update markers

Marker	DS	(s)he said	I said	Context
Oh	170	292	218	17
Well	202	299	502	22
Total	**372**	**591**	**720**	**39**
Total turns	5315	5889	5315	500
% total turns	7.00	10.04	13.55	7.8

Table 4: Frequency of Update Markers

The raw frequencies for the distribution of turn-initial update markers are provided in Table 4. The 'reactive' change of state token 'oh' is more common in reported speech than all direct speech ($\chi^2_{(1)}$ = 16.7, p<0.01) but there is no difference in frequency between self-reported speech and the Local Context turns by the same speaker ($\chi^2_{(1)}$ = 0.58, p=0.45). 'Oh' is however, slightly more frequent in other-reported speech (he/she) than self-reported speech ($\chi^2_{(1)}$ = 4.72, p=0.03).

As Table 4 shows, differences in the use of the ‘prospective’ update marker ‘well’ are more marked. It is approximately twice as common in reported speech as direct speech ($\chi^2_{(1)}$ = 70.9, p<0.01). Most of this difference is accounted for by the use of ‘well’ in self-reported speech where it is approximately twice as common as in the Local Context speech turn by the same speaker ($\chi^2_{(1)}$ = 14.2, p<0.01) and approximately twice as common in self-reported speech than other reported speech ($\chi^2_{(1)}$ = 80.3, p<0.01).

Overall, in contrast to markers of (dis)agreement, signals of updates are more common in reported speech. The use of the reactive ‘oh’ is more strongly associated with other-reported speech whereas the use of the prospective ‘well’ is associated with self-reported speech.

4.4 Contrast and Emphasis

The counts for markers of contrast and emphasis i.e. negations, contrastive conjunctives (but, though), adverbial emphasisers (actually, certainly, clearly, definitely, indeed, obviously, plainly, really, surely, for certain, for sure, of course) and common swearwords are provided in Table 5. For all these markers occurrences at any position within a turn were included for analysis.

Feature	DS	(s)he said	I said	Context
Negation	624	1300	1211	148
Swearwords	6	90	132	3
Contrastives	298	316	411	62
Adverbials	187	162	158	40
Total	**1115**	**1868**	**1912**	**253**
Total turns	5315	5889	5315	500
% total turns	20.98	31.72	35.97	50.6

Table 5: Frequency of Negations and Adverbial emphasises

It is immediately clear from Table 5 that swearwords are much more common in reported speech than in direct speech ($\chi^2_{(1)}$ = 92.5, p<0.01); they are also more common in self-reported speech than other-reported speech ($\chi^2_{(1)}$ = 76.8, p<0.01). Swearwords are also four times more common in self-reported speech than in the Local Context turns by the speaker ($\chi^2_{(1)}$ = 7.15, p<0.01).

Negations follow a similar pattern. They are approximately twice as common in reported speech as direct speech ($\chi^2_{(1)}$ = 266, p<0.01) and approximately twice as common in self-reported speech as other-reported speech ($\chi^2_{(1)}$ = 350, p<0.01). However, negations are less frequent in self-reported speech than in the Local Context turns by the same speaker.

Contrastive conjunctives are also more common in reported speech than direct speech ($\chi^2_{(1)}$ = 4.82, p=0.03) and more than twice as common in self-reported speech than in other-reported speech ($\chi^2_{(1)}$ = 207, p<0.01). However, like negations they are less frequent in self-reported speech than in the Local Context turns by the same speaker ($\chi^2_{(1)}$ = 13.3, p<0.01).

The pattern for adverbial emphasisers is different to the other markers of contrast. Emphasis is both slightly more common in direct speech than reported speech ($\chi^2_{(1)}$ = 5.31, p=0.02) and equally frequent in self-reported and other-reported speech ($\chi^2_{(1)}$ = 0.48, p=0.48). It is also approximately twice as common in the Local Context sample of the speaker (context sample) than in their self-reported speech. Overall, emphasis is slightly more common in direct speech overall and particularly common in turns introducing reported speech.

5 Discussion

Although the results show a clear preference for agreement in direct speech in conversation they also show that, contrary to the predictions of the politeness hypothesis, reported speech does not appear to be a context in which explicit disagreements are more likely to be exposed. On the contrary, people are far less likely to include explicit markers of agreement or disagreement in reported speech than they use directly. Moreover, where they do formulate a reported utterance with an explicit marker it is equally likely to be agreement or disagreement.

Explicit makers of agreement and disagreement are rare of course and not an essential part of actually enacting an agreement or disagreement. However, the results show the same pattern for the less direct markers of agreement and disagreement identified by Walker et al. (2012). Again, markers of both disagreement and agreement are more common in direct speech than reported speech. Overall, it appears that reported speech is not a context in which disagreements are normally re-presented or rehearsed as disagreements.

These results also run counter to the hypothesis that the format of reported speech turns is constrained by concerns with self-presentation. The results are contrary to predictions 5,6 and 7. Although the self-presentation hypothesis predicts that disagreement should not be more common in reported speech, it is incompatible with the observation that it is more common in direct speech and more specifically more common in self-reported speech than other-reported speech. A self-presentation account is also difficult to reconcile with the observation that ostensibly taboo swearwords are more common in direct than reported speech; self or other.

The hypothesis that provides the best fit to the preceding results is Contrastive Stance. The results suggest that reported speech is not used for the re-presentation of (dis)agreements, or at least not in the same way in which they are actually enacted in direct speech. Firstly, the update markers 'Oh' and 'Well' appear to be quite strongly associated with reported speech. This suggests people are deliberately highlighting moments of change more than they actually mark them in direct speech. Although not directly predicted the additional observation that people are more likely to 'well'-preface a self-report of their own remarks and 'oh'-preface reports of another's remark suggests individuals position themselves as delivering updates and report on others receiving them. This asymmetric highlighting of changes in epistemic stance fits with a concern to re-present the newsworthy and contrastive elements of prior conversations. Within these reports what is selected for inclusion also appears to focus on the substance of a dispute, i.e. on expressions of contrast and features that indicate shifts in stance. This is compatible with the relatively low frequency with which 'meta' agreement and disagreement markers are used. It is also compatible with the increased use of use of negations and contrastive conjunctives.

However, there are also some challenges to the Contrastive Stance hypothesis in the data presented above. It doesn't directly account for the observation that swearwords will be used more frequently unless these are also construed primarily as markers of contrast, perhaps acting as and emphasis device. This is plausible but post-hoc. Also, its prediction that markers of emphasis should be more common in reported speech is not borne out. The results show that the turn preceding reported speech (the 'Local Context' turn) does tend to include emphasis so this might reflect a marking of stance but again, this is a post-hoc explanation.

As such, it appears that highlighting points of contrast and representing stance and shifts in assessed parameters are key functions of reported speech. While this study shows that reported speech is not used to re-present how disagreements were enacted, it is possible that other forms of report may. The dataset we worked with predominantly included direct reported speech or quotatives ('he said cats are bad'), but also some indirect reported speech ('he said that cats are bad'). Further work to investigate how the more descriptive indirect reports, and the wider gamut of reported thoughts might be used to re-present disagreement may provide further insights into the reporting of disagreement.

Acknowledgments

This work was funded by EPSRC through the Media and Arts Technology Programme, an RCUK Doctoral Training Centre EP/G03723X/1. The authors would like to thank the reviewers for their insightful comments.

References

Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2012. How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 10–18. Association for Computational Linguistics.

Douglas Biber and Edward Finegan. 1989. Styles of stance in english: Lexical and grammatical marking of evidentiality and affect. *Text-interdisciplinary journal for the study of discourse*, 9(1):93–124.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Lou Burnard. 1995. *British National Corpus: Users Reference Guide British National Corpus Version 1.0*. Oxford Univ. Computing Service.

Herbert H. Clark and Richard J. Gerrig. 1990. Quotations as demonstrations. *Language*, pages 764–805.

Rebecca Clift. 2006. Indexing stance: Reported speech as an interactional evidential1. *Journal of sociolinguistics*, 10(5):569–595.

Rebecca Clift. 2007. Getting there first: non-narrative reported speech in interaction. In E. Holt and R. Clift, editors, *Reporting Talk: Reported Speech in Interaction*, Studies in Interactional Sociolinguistics, chapter 5, pages 120–149. Cambridge University Press.

Susan Conrad and Douglas Biber. 2000. Adverbial marking of stance in speech and writing. *Evaluation in text: Authorial stance and the construction of discourse*, pages 56–73.

Elizabeth Couper-Kuhlen. 2007. Assessing and accounting. In E. Holt and R. Clift, editors, *Reporting Talk: Reported Speech in Interaction*, Studies in Interactional Sociolinguistics, chapter 4, pages 81–119. Cambridge University Press.

Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, pages 97–121.

Markku Haakana. 2007. Reported thought in complaint stories. In E. Holt and R. Clift, editors, *Reporting Talk: Reported Speech in Interaction*, Studies in Interactional Sociolinguistics, chapter 6, pages 150–178. Cambridge University Press.

John Heritage and Steven Clayman. 2010. *Talk in action: Interactions, identities, and institutions*. Wiley. com.

John Heritage. 1984. A change-of-state token and aspects of its sequential placement. *Structures of social action: Studies in conversation analysis*, pages 299–345.

John Heritage. 1998. Oh-prefaced responses to inquiry. *Language in society*, 27(3):291–334.

Elizabeth Holt. 2000. Reporting and reacting: Concurrent responses to reported speech. *Research on Language and Social Interaction*, 33(4):425–454.

Elizabeth Holt. 2007. 'i'm eyeing your chop up mind': reporting and enacting. In E. Holt and R. Clift, editors, *Reporting Talk: Reported Speech in Interaction*, Studies in Interactional Sociolinguistics, chapter 3, pages 47–80. Cambridge University Press.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288.

Helga Kotthoff. 1993. Disagreement and concession in disputes: On the context sensitivity of preference structures. *Language in Society-London-*, 22:193–193.

Anita Pomerantz. 1977. *Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes*. Centre for Socio-Legal Studies.

Anita Pomerantz. 1984. Giving a source or basis: The practice in conversation of telling 'how i know'. *Journal of pragmatics*, 8(5):607–625.

Matthew Purver. 2001. SCoRE: A tool for searching the BNC. Technical Report TR-01-07, Department of Computer Science, King's College London, October.

Randolph Quirk and David Crystal. 1985. *A comprehensive grammar of the English language*, volume 6. Cambridge Univ Press.

Harvey Sacks and Gail Jefferson. 1995. *Lectures on Conversation: Volumes I & II*. Number v. 1-2 in Lectures on conversation / Harvey Sacks. Ed. by Gail Jefferson. Blackwell.

Harvey Sacks. 1987. On the preferences for agreement and contiguity in sequences in conversation. *Talk and social organization*, 54:69.

Emanuel A. Schegloff and Gene H. Lerner. 2009. Beginning to respond: Well-prefaced responses to wh-questions. *Research on language and social interaction*, 42(2):91–115.

Deborah Schiffrin. 1988. *Discourse Markers*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Suzanne Scott. 2002. Linguistic feature variation within disagreements: An empirical investigation. *Text*, 22(2):301–328.

Diane Vincent and Laurent Perrin. 1999. On the narrative vs non-narrative functions of reported speech: A socio-pragmatic study. *Journal of Sociolinguistics*, 3(3):291–313.

Marilyn A. Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Shifting Opinions: An Experiment on Agreement and Disagreement in Dialogue

Shauna Concannon **Patrick G. T. Healey** **Matthew Purver**
Queen Mary University of London
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
`s.concannon@qmul.ac.uk`

Abstract

Disagreement is understood to be socially problematic; it also rarely surfaces in naturally occurring conversation. An experiment was designed to allow us to directly manipulate the occurrence of exposed (dis)agreement and track its effects on the subsequent dialogue. This is the first experiment to directly manipulate the occurrence of exposed agreement and disagreement in dialogue. Insertions of exposed disagreement disrupt dialogues, bringing the topic of disagreement directly into the conversation, provoking clarification requests and resulting in a greater number of self-edits when formulating turns. The insertion of disagreement also led to more instances of exposed agreement, suggesting that dialogue partners co-operate to redress the face-threat of disagreement. Conversely, exposed agreement insertions were not as incongruous and had less disruptive impact on the ensuing dialogues; however, introducing agreement into the dialogue did lead to greater deliberation, with more alternative scenarios considered by participants during the task.

1 Introduction

Disagreeing or expressing a view in opposition to that of your interlocutor can be socially problematic. Disagreement has been associated with confrontation and conflict. Brown and Levinson (1987), in their seminal work on politeness, explain the predisposition for the avoidance of disagreement in terms of *face*, the concept derived from Goffman, relating to the public self-image or identity of an individual in interaction with others (Goffman, 1967). Direct challenges to a speaker or disagreeing with their assertion in dialogue can constitute, in Brown and Levinson's terminology, what is known as a *Face Threatening Act*, that is to say it can threaten the hearer's public self-image.

Conversation Analysts have shown that when people produce assessments of situations or events, positive responses are made more quickly and clearly than negative or unaligned responses (Sacks, 1987; Pomerantz, 1977). Negative or *dispreferred* responses are normally produced more slowly and are often prefaced with some form of agreement (e.g. 'Oh yes... but') and the negative assessment itself is often delayed by several turns and produced with some sort of mitigating account (Pomerantz, 1977). Disagreement, especially when done in a direct manner, is rare in conversation (Concannon et al., 2015). This means that it is difficult to assess what effects it has upon a dialogue. An experimental approach has the advantage that it allows us to directly manipulate the occurrence of exposed (dis)agreement and track its effects on the subsequent dialogue.

Previous studies on disagreement take a distributional or corpus based approach at evidencing and analysing instances of disagreement in interaction (Walker et al., 2012; Abbott et al., 2011; Misra and Walker, 2013; Holtgraves, 1997). These studies have provided valuable insights into the ways in which these complex social interactions are handled in different contexts, and given rise to various theories on how we process, respond to and mitigate the impact of disagreement. However, the literature also highlights that exposed disagreement rarely surfaces in naturally occurring conversation (Pomerantz, 1977; Concannon et al., 2015).

This paper outlines an experimental approach for investigating disagreement, which provides opportunity to manipulate the occurrence of *exposed* (dis)agreement in dialogue. By *exposed*, we refer particularly to direct and unequivocal presentations of agreement and disagreement, such as ‘I agree’ and ‘You’re wrong’. However, we also explore less direct markers, which can, but do not always function in a (dis)agreement capacity. For example, turn initial ‘no’ and ‘yes’, can and are often used to signal agreement and disagreement, however, the function of these markers is context specific and depends on the preceding content (for example a ‘no’ following a negative statement can function as agreement).

1.1 Politeness and Accommodation Theory

One argument for the scarcity of disagreement in dialogue is anchored to the concept of politeness. Politeness Theory builds upon Ervin Goffman’s concept of *face*. Goffman (1967) defines face as ‘the positive social value a person effectively claims for himself’ through interaction and offers a model of co-operation that is enacted when an individual’s face or social value is threatened during interaction. Goffman stresses the co-operative nature of facework: ‘When a face has been threatened ... lack of effort on the part of one person induces compensative effort from others’ (Goffman, 1967). This mutual co-operation and shared consideration in interaction has also been located as a central notion for Politeness Theorists (Brown and Levinson, 1987; Watts, 2003).

Politeness Theory suggests that interlocutors minimise disagreement to save face, employing strategic conflict avoidance techniques to mitigate the effect of any disagreement that may surface (Leech, 1980). However, Accommodation Theory would posit that if someone is agreeable their conversational partner would match them in this convivial approach, whereas if they are adopting a discursive or even combative linguistic style, then their conversational partner would be likely to adopt a similar tact and synchronicity would become more exaggerated (Giles and Smith, 1979). Accommodation Theory posits that interlocutors adopt strategies of *convergence* to integrate and identify socially with another (Giles et al., 1991); this involves the adoption of linguistic similarities and leads to perceived communicative effectiveness (Giles and Smith, 1979) and cooperativeness (Feldman, 1968). Conversely, speech *divergence* reflects distancing from the co-conversant and can surface when confronted with perceived differences to the co-conversant.

1.2 Polite disagreement: When the context is right

Recent literature on disagreement and politeness theory in Sociolinguistics and Conversation Analysis suggests that in certain contexts disagreement is appropriate (Kotthoff, 1993), can signal sociability and intimacy (Schiffrin, 1984; Tannen, 1984; Angouri and Tseliga, 2010), and rather than lead to conflict, help strengthen relationships (Georgakopoulou, 2001; Sifianou, 2012). Furthermore, Chiu (2008) found in problem solving dialogues that disagreement, *when done politely*, was more productive in provoking novel contributions from participants than agreement. So although disagreement, particularly when executed impolitely, tends to be problematic, for certain contexts, such as problem solving and discussion tasks, it may be essential in advancing the deliberative quality of a dialogue. Chiu (2008) also suggests that *agreement* can be potentially detrimental to a dialogue, but the problematic aspects of agreement are not well reported in the literature; this gives rise to the question, ‘what effect does exposed *agreement* have upon a dialogue?’ If it is problematic, how and in what ways does this manifest?

If disagreement encourages novel contributions does agreement, conversely, stifle them? If people are too readily agreeing, does this prevent more involved discussion that could lead to shifts in stance or the development of new contributions? In order to understand the effects of both exposed agreement and disagreement, an experiment was designed that enabled the manipulation of such features under controlled conditions.

1.2.1 Can disagreement lead to more considered discussion?

A motivating factor behind this research is an interest in how individuals are led to shifts in stance, and how and when this occurs through interaction. Although there is good reason to think that disagreement ought to be socially problematic, as well as the insights provided by Chiu (2008), research on the phenomenon of repair shows that disruption in interaction can also be potentially beneficial to the progression of a dialogue (Healey, 2008; Colman et al., 2011), particularly if focused on the clarification of a content issue. Although instances of repair seemingly interrupt the flow of a dialogue, this attempt to address problematic talk is not necessarily negative, rather it seems to drive the conversation forward. Issuing only agreements can often lead to a lack of mutual intelligibility in fact, which is why instances of repair are so common in task-oriented dialogues (Colman et al., 2011), a context where effective coordination is critical to the interactional outcome. Healey (2008) demonstrates that repair processes deal directly with misalignments and have a positive effect on measures of interactional outcome. Consequently, disagreement ought to be a catalyst or precursor to a potential shift in stance, as it signals a direct challenge to a held idea, which in turn may be retained, re-negotiated or or more fundamentally re-conceived. This, together with the findings by Chiu (2008), suggests that disagreement can play an important role in the deliberation and problem solving process.

1.3 Predictions

Given the literature we would expect that exposed disagreement would be especially problematic; it should instigate additional work being done in the interaction and more instances of repair. Insertions of exposed disagreement should be more disruptive than exposed agreement insertions, which should in turn facilitate more agreement. Assuming speakers are being co-operative, all things being equal, then disagreement should lead to more hedging and mitigation in order to manage the disagreement and minimise face threat. However, it may also lead to additional shifts in stance, or the consideration of more alternatives during the discussion dialogues.

2 Agreement and Disagreement Fragment Experiment

In order to assess the impact of exposed (dis)agreement, an experiment was designed in which instances of exposed (dis)agreement were artificially inserted into a dialogue. Turn-initial discourse markers such as 'No', 'But' 'You're wrong' and, 'I disagree' can highlight instances of disagreement within a conversation. Similarly, 'Yes', 'And', 'I agree' and 'You're right' can serve as indicators of agreement, or reinforce congruence. These eight fragments were selected because they provide a range of exposed, direct (dis)agreement and more subtle markers that *can* be used in (dis)agreement.

2.1 Hypotheses

1. Accommodation Theory: The general accommodation hypothesis is that dialogue partners match linguistic and discursive style. Thus the general *accommodation hypothesis* predicts that the insertion of agreement fragments will elicit additional instances of agreement, while the insertion of disagreement fragments will elicit additional instances of disagreement.

2. Politeness: The general politeness hypothesis is that face-threatening acts are socially problematic and should result in compensatory action being taken to redress and mitigate the situation. The general *politeness hypothesis* thus predicts that inserting disagreement fragments into a dialogue should lead to more work being done and more cooperation and consideration being displayed; this may result in increased effort when formulating responses (higher number of self-edits) and more clarification requests, expressions of agreement and other routinised polite sequences.

3. Productive Disagreements: The general productive disagreement hypothesis is that disagreement is essential for advancing the deliberative quality and problem solving aspects of dialogue. The *productive disagreement hypothesis* thus predicts that people will respond constructively to disagreement. The specific predictions for particular response measures are a much lower level issue, but we would expect the insertion of disagreement fragments to lead to increased deliberation taking place which lead to a higher number of shifts in stance over the course of a dialogue.

2.2 Method

Pairs of participants were seated at separate computers in adjacent rooms and given an instruction sheet to read detailing the balloon task. Participants are presented with a fictional scenario in which an hot air balloon is losing altitude and about to crash. The only way for any of three passengers to survive is for one of them to jump to a certain death. The three passengers are: Dr. Nick Riviera, a cancer scientist, Mrs. Susie Derkins, a pregnant primary school teacher, and Mr. Tom

Derkins, the balloon pilot and Susies husband. Participants are told to take as much time as they need to read the summary of the situation and then discuss with their partners via a chat tool set up on the computer at which they are seated, and attempt to come to a conclusion over who should jump from the balloon. The advantages of this task are that it is effective at generating debates between subjects and involves articulations of agreement and disagreement as they attempt to come to a conclusion. There is also plenty of scope for deliberation and shifts in stance.

2.3 Participants

Seventy-two participants were recruited, 46 female and 26 male, with the majority being undergraduate and postgraduate students at the University of London. Participants were invited to attend with someone who they already knew. They were recruited in pairs to ensure that inter-pair participants were acquainted. For a couple of experiments if one participant didn't show up a stand in was recruited last minute, and in these exceptions, which are marked in the data, the pair were not previously acquainted with each other. Each participant was paid at a rate of £7.50 per hour for participating in the experiment, or if they were a Psychology student at Queen Mary University of London then they could receive course credits in lieu of payment.

2.4 Materials

The participants communicate via a specially programmed chat tool, similar to other instant messenger interfaces they may have used previously. The Dialogue Experimental Toolkit (DiET) chat tool is a text-based chat interface facilitating real time manipulations of the dialogue. It is possible to programme several different types of interventions using the chat tool: turns may be altered prior to transmission, turns may not be relayed, and additional turns may be added, (e.g. Healey et al. (2003), insertion of spoof clarification requests).

These manipulations occur as the dialogue progresses, thus making them minimally disruptive to the sequence of dialogue. The DiET chat tool is built in Java and consists of a server console and user interface. Participants are faced with a text box displaying the conversation history and a smaller text box into which they can type. Participants can type simultaneously and their message is relayed to their conversation partner by use of the ENTER key. The server time stamps and stores all key presses. All turns are passed to the server before being transmitted to the other participant, thus making it an intermediary between what the participants type and what they receive. Turns can be automatically altered, removed or inserted by the server before they are relayed.

2.5 Design

The experiment is conducted in pairs; there were 12 dyads for each condition. Pairs of participants were presented with a discussion task and instructed to discuss for 30 minutes and attempt to come to an agreement. Each pair of participants was assigned to a condition at random. There were three experimental conditions. Please note, what we gloss here as the *Agreement* and *Disagreement* conditions, are named as such because the inserted fragments in each condition *can* index disagreement, however, we recognise that the more indirect fragments do not consistently perform this function.

- **Control condition**: Participants are welcomed and briefed before being sat at their respective computers, which were situated in adjoining rooms. They receive their task instructions on a piece of paper and can start when they are ready. They are instructed to discuss the scenario and attempt to come to an agreement on who should jump from the balloon for 30 minutes. No interventions are performed by the server; participants receive the dialogue turns exactly as they were typed.

- **Agreement condition**: Initial procedure is exactly the same as the control condition. Participants receive the dialogue turns exactly as they were typed, except for every fourth turn when one of the following fragments is inserted position: *you're right, I agree, yes, and.*

- **Disagreement condition**: Initial procedure is exactly the same as the control condition. Participants receive the dialogue turns exactly as they were typed, except for every fourth turn when one of the following fragments inserted at turn-initial position: *you're wrong, I disagree, no, but.*

A small scale pilot study raised some design implications that were accordingly addressed. In the

Agreement and Disagreement conditions manipulations were carried out every fourth turn issued by each speaker as this was deemed an acceptable frequency for interventions without proving too disruptive to the conversation. No intervention was made if the turn consisted of only one word, or the turn started with the same text as featured in the insertion fragments. This was to avoid the production of particularly non-sensical turns such as *you're wrong I agree*. The fragments were cycled through in order but the exposed (dis)agreement fragments (*you're wrong/right, I (dis)agree*) appeared half as often due to their marked nature.

3 Results

Data was gathered both directly from the chat tool which logged various features such as typing time, number of self-edits, i.e. use of the backspace and delete key and temporal data, as well as the transcripts themselves, which were analysed for linguistic features and frequencies. Additionally the resulting transcripts were hand coded for clarification requests and *stance shifts*, explained in more detail below.

3.1 A note on terminology

Turn: For the purpose of this experiment, a turn constitutes the text relayed in a single message, meaning what is delineated by the ENTER key.

Intervention Turn (IT): The IT refers to the turn issued by a speaker which has had a Turn-initial intervention fragment inserted before the actual typed message.

Intervention Reply Turn (IRT): The IRT refers to the next turn issued by the speaker who receives the Intervention Turn. This is not always the next sequential turn after the IT, as the speaker whose turn contained the IT may issues another turn.

Clarification Requests: The transcripts were hand coded for Clarification Requests (CR), a form of repair in which speakers signal a need for further information, typically due to a lack of full comprehension of a previous utterance. This was done by a single annotator, blind, and all labelling indicating which condition a file belonged to was removed. CRs were hand labeled in the dataset, based on Purver et al. (2003) schema, example provided in Table 1 .

Stance shifts: The transcripts were hand coded for shifts in stance regarding who to throw off of the balloon, i.e when a participant changed their point of view over who to sacrifice or save. There were seven potential stance states that cover all the possible combinations of who to save and who to sacrifice[1]. This was done by a single annotator, blind, and all labelling indicating which condition a file belonged to was removed. A participant's stance was carried over to the next turn, unless it provided new information that contradicted the previous stance.

Turn 1:	P1	**you're wrong** or maybe we are just going by gender stereotypes.. the feminist in me is screaming	IT
Turn 2:	P1	haha	
Turn 3:	P2	what if thats the whole point	IRT
Turn 4:	P1	sorry what if....?	CR
Turn 5:	P1	susie jumped?	CR

Table 1: Example of Reply Turn labelling

3.2 Overview of dataset

Table 2 displays the descriptive data for the turn, word and character counts for each condition.

	Condition		
Avg.	Control	Agreement	Disagreement
Turns by Dyad	86.71	63.17	70.79
Words by Dyad	587.67	555.58	535.08
Char. by Dyad	2938	2797	2710
Words per turn	7.41	9.49	9.11

Table 2: Summary of average typed data per condition

Both intervention conditions result in fewer overall turns than in the Control condition, but this was particularly the case, and statistically significant, with the Agreement condition (positive and agreement insertions, such as *yes* and *I agree*). Although the Agreement condition features fewer

---

[1] The range of possible stances: 1. Undecided, 2. Save Susie but undecided on who should die, 3. Save Nick but undecided on who should die, 4. Save Tom but but undecided on who should die, 5. Sacrifice Susie (and therefore save the other two), 6. Sacrifice Nick, 7. Sacrifice Tom.

turns than the Control condition, there are more words per turn on average in the Agreement condition. A non-parametric Kruskal Wallis test confirms a significant overall effect of Condition on the turns typed in the dialogues ($H_{(2)}$ = 6.34, p<0.04).[2] Subsequent planned pairwise comparisons with the Dunns test showed a significant increase in the number of turns per dyad in the Control condition compared to the Agreement condition (p<0.05). There is an overall effect of condition on the distribution of average words per turn, as confirmed by a non-parametric, Kruskal Wallis test ($H_{(2)}$ = 6.55, p<0.04). Subsequent planned pairwise comparisons with the Dunns test showed a significant increase between Agreement and Control conditions (p<0.03).

3.3 Message construction

Condition	Typing Time	Self-edits
Control	11850	6.98
Agree	16210	6.97
Disagree	13484	7.51

Table 3: Table depicting averageTyping Time and number of Self-edits (delete key presses), per turn, per condition

Table 3 shows the average typing time in milliseconds and the number of self-edits per turn. Self-edits are represented by the number of times the delete key is pressed during turn construction. A non-parametric Kruskal Wallis test finds an omnibus effect of condition on the number of self-edits during turn construction ($H_{(2)}$ = 40.92, p<0.01), with planned pairwise comparison revealing significant difference between the Agreement and Disagreement conditions (p<0.01). An overall effect of condition on typing time is confirmed by a non-parametric Kruskal Wallis test ($H_{(2)}$ = 99.28, p<0.01), with planned pairwise comparison revealing significant difference between the Agreement and Control conditions (p<0.01).

3.4 Message content

The following tables highlight differences in the content of the dialogues, such as Clarification Requests and instances of exposed and potential disagreement.

3.4.1 Clarification Requests

Condition	Total Number of CRs	Mean CRs per dyad
Control	10	0.42
Agreement	13	0.54
Disagreement	50	2.08

Table 4: No. of Clarification requests by Condition

Table 4 shows the number of Clarification Requests by condition. The Disagreement condition has a significantly higher number of Clarification Requests than Control condition and Agreement condition. A non-parametric Kruskal Wallis test confirms an overall effect of Condition on the number of Clarification Requests in the dialogues ($H_{(2)}$ = 12.03, p<0.01). Planned pairwise comparison showed a significant increase between Control and Disagree conditions (p<0.01) and Agree and Disagree (p<0.02).

3.4.2 Instances of exposed and potential (dis)agreement

Table 5 shows the frequencies of turn-initial exposed and potential (dis)agreement markers. The markers included here are the same ones that feature in the fragments that were artificially inserted during the experiment.

Turn-initial	Control condition	Agreement condition	Disagreement condition
Exposed (dis)agreement			
I agree	2	5	10
You're right	0	0	0
I disagree	0	0	3
You're wrong	0	0	3
Totals:	2	5	16
Yes	170	124	139
No	29	23	35
And 103		51	55
But	119	81	77

Table 5: Table providing frequency data of turn-initial content of messages relayed during experiment dialogues.

---

[2] Throughout we use p<0.05 as our criterion level but report computed probabilities to two decimal places for completeness.

Exposed (dis)agreement is more frequent in the Disagreement condition. A non-parametric Kruskal Wallis test shows a significant omnibus effect of condition on turn-initial exposed (dis)agreement ($H_{(2)}$ = 9.74, p<0.01). Subsequent planned pairwise comparisons with the Dunns test showed a significant increase in the number of instances of exposed (dis)agreement in the Disagreement condition compared to the control condition (p<0.01).

3.5 Deliberation and shifts in stance

The experiment transcripts were also hand labeled for *stance shifts*, i.e. when a participant voices a departure from one held opinion to an alternative regarding who should jump from the balloon.

Condition	Total	Median	Mean	St. Dev.
Control	175	7.5	8.33	3.96
Agree	248	11	10.33	4.88
Disagree	175	6	7.29	4.31

Table 6: Total number of stance state changes and averages per participant

The total number of state changes and average per participant by condition are shown in Table 6. The median number of stance state changes per participant is significantly effected by condition ($\chi(2)$ = 6.91, p=0.03). A Median Test was conducted as the variance is not approximately equal across samples, being much larger for the agreement condition. This result suggests that the Disagreement condition tends to reduce the number of alternatives people will consider and the agreement condition tends to increase it.

There is no correlation between the length of the conversation (in turns) and the number of state changes (Kendals Tau = -0.007, p = 0.94), so the significance is not related to nor skewed by the fact that the Agreement condition contains longer dialogues, i.e. it is not just about how much participants talk.

4 Discussion

The turn-initial frequency data shows that exposed agreement *and* disagreement are more common the Disagreement condition. The is counter to our Accommodation hypothesis, which anticipated that agreement would lead to more agreement while disagreement would engender more disagreement. Although there are notably zero instances of exposed disagreement in the Agreement condition, the comparative frequency in the Disagreement condition did not confirm a significant effect of condition. Furthermore, a third of the instances of turn-initial exposed disagreement in the Disagreement condition are actually instances of repair, rather than disagreement. As shown in the following excerpt from an experiment transcript, the exposed disagreement is incongruent, jarring and provokes a repair sequence. The respondent quotes back the source of trouble, indicated by the asterisks in the example below, which were falsely counted as turn-initial disagreement. The artificial insertions are shown in square brackets:

> *Example 1* A: Pros of keeping the doctor alive
> A: [you're wrong] cures cancer
> B: [no] you're wrong?*
> A: What about?
> B: no, I don't understand what you just said
> B: You're wrong cures cancer?*
> A: The doctor, if still alive will be about to discover the sure for the 'most common types of cancer'

This example demonstrates the disruptive nature of the inserted disagreement fragment; it disrupts the dialogue and is deemed incongruous enough for participant B to comment on, while participant A simply carries on with the conversation. This occurred several times in the dataset, however, only ever with the exposed disagreement fragments and never with the exposed agreement fragments. In line with the literature we found that exposed disagreement is especially problematic and on one occasion the insertion was so problematic that it was directly referenced and quoted by a participant, with both participants being alerted to the intervention.

> *Example 2* A: imagine how many scientists in the world
> B: you're wrong theres a lot
> A: i'm wrong?
> B: what?
> A: you said 'you're wrong theres a lot'
> A: [no] what am i wrong about?

The Disagreement condition featured a significantly higher number of clarification requests. The

Productive Disagreement hypothesis anticipated that this would signal additional work being done by participants trying to more fully understand one another's point of view. However, it is possible that the clarification requests are more clausal clarification than an attempt to understand the content; this interpretation is supported by the Example 2, which notably features a high number of clarification requests in a very short segment of dialogue, however, further analysis is needed to confirm this.

The Productive Disagreement hypothesis also anticipated that the Disagreement condition would lead to more *stance states* bang considered. The results show that although there is an effect of condition on the number of different stance states or scenarios considered during the dialogue, the directionality was contrary to our predictions. The insertion of agreement fragments led to more shifts in stance. This may be due to the particularly marked and direct nature of the exposed disagreement fragments, which may have closed the discussion down. This would align with the CA and Politeness Theory literature, as well as Chiu (2008), which specifies that while *polite* disagreement may yield more novel contributions, impolite disagreement is always problematic.

Overall, our results most strongly confirm the Politeness hypothesis. Insertions of exposed disagreement had a disruptive effect upon the dialogues, producing confusion and clarification requests due to their unexpected and incongruous nature. Conversely, exposed agreement, even though also inserted randomly, did not disrupt the dialogue in the same way and were never explicitly addressed by a participant. The Disagreement condition produced significantly more instances of exposed agreement, which is most easily interpreted in terms of politeness, face and redressive action; with additional exposed disagreement being introduced into the dialogues, it seems that participants respond with cooperation and attempt to redress the potential affronts to face posed by the inserted fragments. As predicted there were more self-edits in the Disagreement condition, suggesting that participants were having to work harder to respond to the potentially face threatening insertions. Our results most strongly support the Politeness hypothesis and confirm that exposed disagreement is problematic and disruptive in dialogue.

Acknowledgments

This work was funded by EPSRC through the Media and Arts Technology Programme, an RCUK Doctoral Training Centre EP/G03723X/1. The authors would like to thank the reviewers for their helpful comments.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, pages 2–11. Association for Computational Linguistics.

Jo Angouri and Theodora Tseliga. 2010. "you have no idea what you are talking about!" from e-disagreement to e-impoliteness in two online fora. *Journal of Politeness Research. Language, Behaviour, Culture*, 6(1):57–82.

P. Brown and S.C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Ming Ming Chiu. 2008. Flowing toward correct contributions during group problem solving: A statistical discourse analysis. *The Journal of the Learning Sciences*, 17(3):415–463.

Marcus Colman, Patrick GT Healey, et al. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd annual meeting of the cognitive science society*, pages 1563–1568.

Shauna Concannon, Patrick GT Healey, and Matthew Purver. 2015. Taking a stance: a corpus study of reported speech. In *Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (GoDial)*, Semdial Workshop, page In Press.

Roy E Feldman. 1968. Response to compatriot and foreigner who seek assistance. *Journal of Personality and Social Psychology*, 10(3):202.

Alexandra Georgakopoulou. 2001. Arguing about the future: On indirect disagreements in conversations. *Journal of Pragmatics*, 33(12):1881–1900.

Howard Giles and Philip Smith. 1979. *Accommodation theory: Optimal levels of convergence*.

Howard Giles, Justine Coupland, and Nikolas Coupland. 1991. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Erving Goffman. 1967. On face-work. *Interaction ritual*, pages 5–45.

Patrick Healey. 2008. Interactive misalignment: The role of repair in the development of group sublanguages. *Language in Flux. College Publications*, 212.

Thomas Holtgraves. 1997. Yes, but... positive politeness in conversation arguments. *Journal of Language and Social Psychology*, 16(2):222–239.

Helga Kotthoff. 1993. Disagreement and concession in disputes: On the context sensitivity of preference structures. *Language in Society*, 22:193–193.

Geoffrey N Leech. 1980. *Explorations in semantics and pragmatics*. John Benjamins Publishing.

Amita Misra and Marilyn A Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Conference of the Special Interest Group on Discourse and Dialogue*, page 920.

Anita Pomerantz. 1977. *Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes*. Centre for Socio-Legal Studies.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer.

Harvey Sacks. 1987. On the preferences for agreement and contiguity in sequences in conversation. *Talk and social organization*, 54:69.

Deborah Schiffrin. 1984. Jewish argument as sociability. *Language in society*, 13(03):311–335.

Maria Sifianou. 2012. Disagreements, face and politeness. *Journal of Pragmatics*, 44(12):1554–1564.

Deborah Tannen. 1984. The pragmatics of cross-cultural communication. *Applied Linguistics*, 5(3):189–195.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Richard J Watts. 2003. *Politeness*. Cambridge University Press.

Changing perspective: Local alignment of reference frames in dialogue

Simon Dobnik and Christine Howes
Centre for Language Technology
University of Gothenburg, Sweden
{simon.dobnik,christine.howes}@gu.se

John D. Kelleher
School of Computing
Dublin Institute of Technology, Ireland
john.d.kelleher@dit.ie

Abstract

In this paper we examine how people negotiate, interpret and repair the frame of reference (FoR) in free dialogues discussing spatial scenes. We describe a pilot study in which participants are given different perspectives of the same scene and asked to locate several objects that are only shown on one of their pictures. This task requires participants to coordinate on FoR in order to identify the missing objects. Preliminary results indicate that conversational participants align locally on FoR but do not converge on a global frame of reference. Misunderstandings lead to clarification sequences in which participants shift the FoR. These findings have implications for situated dialogue systems.

1 Introduction

Directional spatial descriptions such as "to the left of green cup" or "in front of the blue one" require the specification of a frame of reference (FoR) in which the spatial regions "left" and "front" are projected, for example "from where I stand" or "from Katie's point of view". The spatial reference frame can be modelled as a set of three orthogonal axes fixed at some origin (the location of the landmark object) and oriented in a direction determined by the viewpoint (Maillat, 2003).

A good grasp of spatial language is crucial for interactive embodied situated agents or robots which will engage in conversations involving such descriptions. These agents have to build representations of their perceptual environment and connect their interpretations to shared meanings in the common ground (Clark, 1996) through interaction with their human dialogue partners. There are two main challenges surrounding the computational modelling of FoR. Firstly, there are several ways in which the viewpoint may be assigned. If the FoR is assigned by the reference object of the description itself ("green cup" in the first example above) then we talk about *intrinsic* reference frame (after (Levinson, 2003)). Alternatively, the viewpoint can be any conversational participant or object in the scene that has an identifiable front and back in which case we talk about a *relative* FoR. Finally, one can also to refer to the location of objects where the viewpoint is external to the scene, for example, as a superimposed grid structure on a table top with cells such as A1 and B4. In this case it is an *extrinsic* reference frame. There are a number of factors that affect the choice of FoR, including: task (Tversky, 1991), personal style (Levelt, 1982), arrangement of the scene and the position of the agent (Taylor and Tversky, 1996; Carlson-Radvansky and Logan, 1997; Kelleher and Costello, 2009; Li et al., 2011), the presence of a social partner (Duran et al., 2011), the communicative role and knowledge of information (Schober, 1995). The second challenge for computational modelling is that the viewpoint may not be overtly specified and must be recovered from the linguistic or perceptual context. Such underspecification may lead to situations where conversational partners fail to accommodate the same FoR leading to miscommunication.

Psycholinguistic research suggests that interlocutors in a dialogue align their utterances at several levels of representation (Pickering and Garrod, 2004), including their spatial representations (Watson et al., 2004). However, as with syntactic priming (Branigan et al., 2000), the evidence comes from controlled experiments with a confederate and single prime-target pairs of pictures, and this leaves open the question of how well such effects scale up to longer unconfined free dialogues. In the case of syntactic priming, corpus studies suggest that interlocutors actually diverge syntactically in free dialogue (Healey et al., 2014).

Semantic coordination has been studied using the Maze Game (Garrod and Anderson, 1987), a task in which interlocutors must produce location descriptions, which can be figurative or abstract. Evidence suggests that dyads converge on more abstract representations, although this is not explicitly negotiated. Additionally, the introduction of clarification requests decreases convergence, suggesting that mutual understanding, and how misunderstandings are resolved is key to shifts in description types (Mills and Healey, 2006). However, both participants see the maze from the same perspective, in contrast to our egocentric, embodied perceptions of everyday scenes.

We are interested in how participants align their spatial representations in free dialogue when they perceive a scene from different perspectives. If the interactive alignment model is correct, although participants may start using different FoRs (using e.g. an egocentric perspective (Keysar, 2007)), they should converge on a particular FoR over the course of the dialogue. We are also concerned with how they identify if a misalignment has occurred, and the strategies they use to get back on track in dialogues describing spatial scenes.

In contrast to several previous studies, this paper investigates the coordination of FoR between two conversational participants over an ongoing dialogue. Our hypotheses are that (i) there is no baseline preference for a specific FoR; (ii) participants will align on spatial descriptions over the course of the dialogue; (iii) sequences of misunderstanding will prompt the use of different FoRs.

2 Method

We describe below our pilot experimental set-up in which participants were required to discuss a visual scene in order to identify objects that were missing from one another's views of the scene.

2.1 Task

Using 3D modelling software (Google SketchUp) we designed a virtual scene depicting a table with several mugs of different colours and shapes placed on it. As shown in Figure 1, the scene includes three people on different sides of the table. The people standing at the opposite side of the table were the avatars of the participants (the man = P1 and the woman = P2), and a third person at the side of the table was described to the participants as an observer "Katie".

Each participant was shown the scene from their avatar's point of view (see Figures 2 and 3), and informed that some of the objects on the table were missing from their picture, but visible to their partner. Their joint task was to discover the missing objects from each person's point of view and mark them on the printed sheet of the scene provided. The objects that were hidden from each participants are marked with their ID in Figure 1.

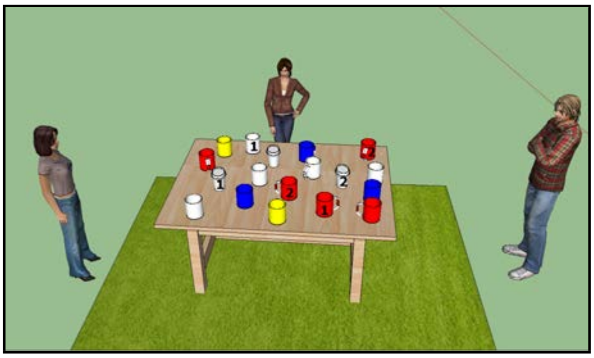


Figure 1: A virtual scene with two dialogue partners and an observer Katie. Objects labelled with a participant ID were removed in that person's view of the scene.

2.2 Procedure

Each participant was seated at their own computer and the participants were separated by a screen so that they could not see each other or each other's computer screens. They could only communicate using an online text based chat tool (Dialogue Experimental Toolkit, DiET, (Healey et al., 2003)).[1] The DiET chat tool resembles common online messaging applications, with a chat window in which participants view the unfolding dialogue and a typing window in which participants can type and correct their contributions before sending them to their interlocutor. The server records each key press and associated timing data.

In addition to the chat interface each participant saw a static image of the scene from their view, as shown in Figure 2, which shows the scene from P1's view and Figure 3, which shows the same scene from P2's view.

2.3 Participants

In the pilot study reported here, we have recorded two dialogues. Both dialogues were conducted in English but the native language of the first pair was Swedish while the second pair were native British English speakers. Participants were instructed that

---

[1] `http://cogsci.eecs.qmul.ac.uk/diet/`

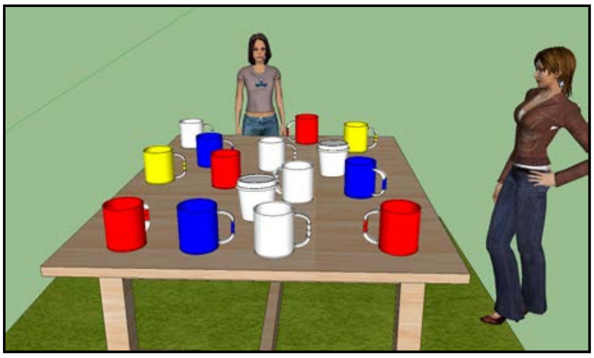


Figure 2: The table scene as seen by Participant 1.

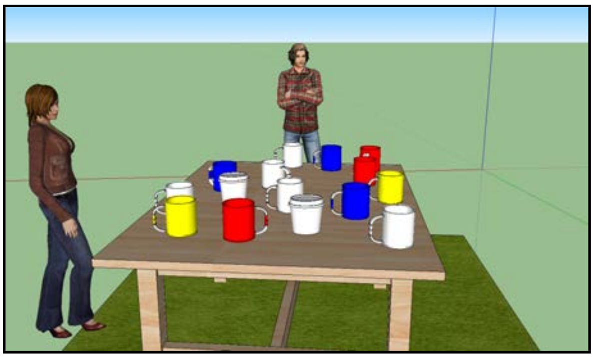


Figure 3: The table scene as seen by Participant 2.

they should chat to each other until they found the missing objects or for 30 minutes. The first dyad took approximately 30 minutes to find the objects and produced 157 turns in total. The second dyad (native English speakers) discussed the task for a little over an hour, during which they produced 441 turns. Following completion of the task participants were debriefed about the nature of the experiment.

2.4 Data annotation

The turns were annotated manually for the following features: (i) does a turn (T) contain a spatial description; (ii) the viewpoint of the FoR that the spatial description uses (P1, P2, Katie, object, extrinsic); a turn may contain several spatial descriptions with different FoR in which case all were marked; (iii) whether a turn contains a topological spatial description such as "near" or "at" which do not require a specification of FoR; and (iv) whether the FoR is explicitly referred to by the description, for example "on my left".

20 *P1:* from her right I see yell, white, blue red
spatial, relative-katie, explicit
21 and the white has a funny thing around the top
22 *P2:* then you probably miss the white i see
23 *P1:* and is between yel and bl but furhter away from katie
spatial, relative-katie, explicit, topological
24 *P2:* because i see a normal mug too, right next to the yellow one, on the left
spatial, relative-katie, topological
25 *P1:* ok, is your white one closer to katie than the yellow and blue?
spatial, relative-katie, topological
26 *P2:* yes
27 closest to me, from right to left:
spatial, relative-p2, topological
28 *P1:* ok, got it
29 *P2:* white mug, white thing with funny top, red mug, yellow mug (the same as katies)

The example also shows that topological spatial descriptions can be used in two ways. They can feature in explicit definitions of FoR as "away" in T23, be independent as "right next to" in T24 and "closest to me" in T27 or sometimes they may be ambiguous between the two as " closer to Katie" in T25. In addition to referring to proximity, topological spatial descriptions also draw attention to a particular part of the scene that dialogue participants should focus on to locate the objects and to a particular FoR that has already been accommodated, in this case relative to Katie. Strictly speaking, this is not an explicit expression of a FoR but is used to add additional salience to it.

2.5 Dialogue Acts and entropy

We tagged both conversations with a dialogue act (DA) tagger trained on the NPS Chat Corpus (Forsyth and Martell, 2007) using utterance words as features as described in Chapter 6 of (Bird et al., 2009) but using Support Vector Machines rather than Naive Bayes classifier (F-score 0.83 tested on 10% held-out data). Out of 15 dialogue acts used, the most frequent classifications of turns in our corpus are (in decreasing frequency) Statement, Accept, yAnswer, ynQuestion and whQuestion and others. In parallel to DA tagging we also marked turns that introduced a change in the FoR assignment. Turns with no projective spatial description and hence no FoR annotation are marked as no-change. We process the dialogues by introducing a moving window of 5 turns and for each window we calculate the entropy of DA assignments and the entropy of FoR changes.

3 Results and Discussion

3.1 Overall usage of FoR

Table 1 summarises the number of turns that use each FoR in the dialogues. The data shows that the majority of FoR is assigned relative to dialogue participants (P1: 36%, P2: 27% and Speaker:

33%, Addressee: 29%, all values relative to the turns containing a spatial description). Extrinsic FoR is also quite common (25%) followed by the FoR relative to Katie (6%). In 10% of turns containing a spatial description the FoR could not be determined, most likely because a turn contained only a topological spatial description. Topological spatial descriptions are used in 18% of spatial turns. Note that since one turn may contain more than one spatial description, the number of turns of these does not add up to the total number of turns containing a spatial description.

Category	Turns	%
Turns in total	598	1.0000
Contains a spatial description	245	0.4097
FoR=P1	88	0.3592
FoR=P2	66	0.2694
FoR=speaker	81	0.3306
FoR=addressee	72	0.2939
FoR=Katie	15	0.0612
FoR=extrinsic	61	0.2490
FoR=unknown	26	0.1061
Topological description	44	0.1796

Table 1: Overall usage of FoR

In our data there are no uses of the intrinsic reference frame relative to the landmark object. This may be because the objects in this study were mugs and they are used as both target and landmark objects in descriptions. Although they may have identifiable fronts and backs and are hence able to set the orientation of the FoR, they are not salient enough to attract the assignment of FoR relative to the presence of the participants. This observation is orthogonal to the observation made in earlier work where the visual salience properties of the dialogue partners and the landmark object were reversed compared to this scene (Dobnik et al., 2014). Note, however, that we annotate descriptions such as “one directly in from of you” (D(ialogue) 1, T146) as relative FoR to P1, although this could also be analysed as an intrinsic FoR. We opt for the relative interpretation on the grounds that otherwise important information about which contextual features attract the assignment of FoR would be lost. In our system there is therefore no objectively intrinsic FoR but FoR assigned to different contextually present entities.

3.2 Local alignment of FoR

Figure 4 show the uses of FoR over the length of the entire D1 and the same length of utterances of D2. The plots show that although there is no global preference for a particular entity to assign the FoR one can observe local alignments of FoR that stretch over several turns which can be observed as lines made of red (P1) and green (P2) shapes. This supports the findings in earlier work (Watson et al., 2004; Dobnik et al., 2014) that participants tend to align to FoR over several turns.

Partial auto-correlations on each binary FoR variable in Figure 4 (P1, P2, Katie and Extrinsic) confirm this. Each correlates positively with itself ($p < 0.05$) at 1–3 turns lag, confirming that the use of a particular FoR makes reuse of that FoR more likely. Cross-correlations between the variables show no such pattern.

The graph also shows that the alignment is persistent to a different degree at different parts of both dialogues. For example, in D1 the participants align considerably in the first part of the dialogue up to turn 75, first relative to Katie, then to P2 and finally to P1. After approximately T115 both FoR relative to P1 and P2 appear to be used interchangeably in a threaded manner as well as the use of the extrinsic perspective. In D2 the situation is reversed. The participants thread the usage of the FoR in the first part of the dialogue but converge to segments with a single FoR shortly before T100 where they both prefer the extrinsic FoR and also FoR relative to P1. We will discuss these segments further in Section 3.4

Overall, the data show that the use of FoR is not random and that different patterns of FoR assignment and coordination are present at different segments of the dialogue. In order to understand how FoR is assigned we therefore have to examine these segments separately.

3.3 Explicitness of FoR

With an increase in (local) alignment, as discussed above, we might expect that there is less necessity for dialogue participants to describe the FoR overtly after local alignment has been established. Explicitness of FoR is therefore indicated in Figure 4: stars indicate that the FoR is described explicitly wheres triangles indicate that it is not. However, contrary to our expectation that the FoR would only be described explicitly at the beginning of a cluster of aligned FoR turns, it appears

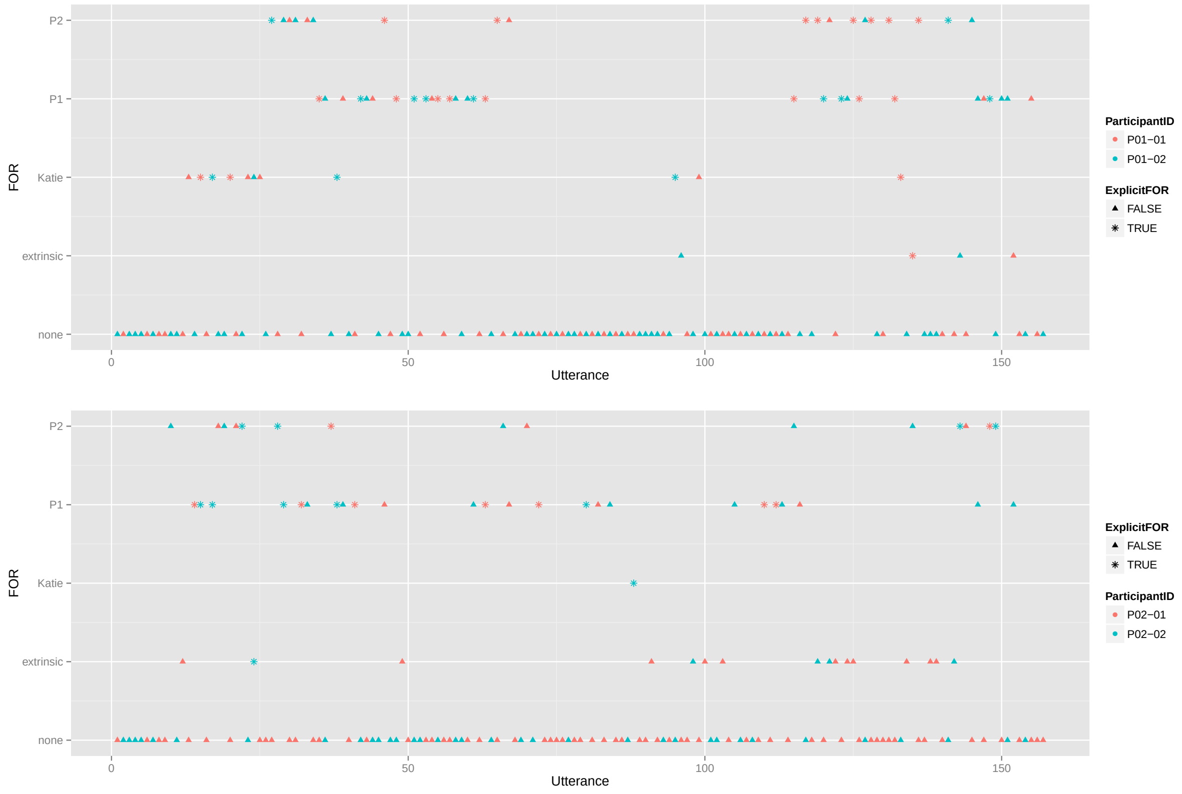


Figure 4: The assignment of FoR over the length of Dialogue 1 (top) and Dialogue 2 (bottom)

that the FoR is explicitly described every couple of utterances even if the participants align as in the first half of D1. This may be because participants are engaged in a task where the potential for referential ambiguity is high and precision is critical for successful completion of the task.

Note also that in D2 at around turn 100 there are clusters of turns where extrinsic FoR was used but this was not referred to explicitly. This is because participants in this dialogue previously agreed on a 2-dimensional coordinate system involving letters and numbers that they superimposed over the surface of the table. Referring to a region “A2” does not require stating “of the table” and hence a lack of explicitness in their FoRs.

3.4 Changing FoR

One of the main consequences of the local, and not global alignment of FoR, as shown in Figure 4 is that there are several shifts in FoR as the dialogue progresses. Below we outline some possible reasons for this, with illustrative examples taken from the dialogues. Due to the sparsity of data in our pilot study, these observations are necessarily qualitative, but they point the way towards some interesting future work.

(i) The scene is better describable from another perspective. Due to the nature of the task and the scene, it is not possible to generate a unique and successfully identifiable referring expression without leading to miscommunication. In D1 we can observe that the dialogue partners take neutral Katie’s viewpoint over several turns. In fact, they explicitly negotiate that they should take this FoR: T13 “shall we take it from katies point of view?”. However, in T25 P1 says “ok, is your white one closer to katie than the yellow and blue?” which prompts P2 to switch FoR to themselves “closest to me, from right to left:”. The change appears to be initiated by the fact that the participants have just discovered a missing white mug but a precise reference is made ambiguous because of another white distractor mug nearby. P2 explicitly changes the FoR because a description can be made more precise from their perspective: from Katie’s perspective both white mugs are arranged in a line at her front. Interestingly, in T35 P1 uses the same game strategy and switches the FoR to theirs saying “closest to me, from left to right red, blue, white, red” and the conversation continues using that FoR for a while, until turn 63. The example also shows that participants align in terms of conversational games for the purposes of identifying the current object and that the nature of dialogue game also affects the assignment of FoR.

(ii) Current dialogue game. The nature of the task seems to naturally lead to a series of different dialogue games, from describing the whole scene to zooming in on a particular area when a potential mismatch is identified. In this case, since the scene in focus is only a part of the overall picture it is less likely that a an identifiable reference to a particular object will fail as there will be fewer distractors. As a result a single FoR can be used over a stretch of the conversation and participants are likely to align. There is less need for explicit perspective marking. See for example D1,T20-29 in the previous dialogue listing which corresponds to a cluster in Figure 4. Another cluster in Figure 4 starts at D1,T42 and is shown below. P2 identifies an empty space in their view which they assume is not empty for P1 and this becomes a region of focus. Since this region is more visually accessible to P1 and since they are information giver they opt for P1's FoR ("away from you" in T42 and T43). As shown in Figure 4 this is a dominant FoR for this stretch of dialogue.

42 *P2:* there is an empty space on the table on the second row away from you
relative-p1, explicit, topological
43 between the red and white mug (from left to right)
relative-p1
44 *P1:* I have one thing there, a white funny top
relative-p1
45 *P2:* ok, i'll mark it.
46 *P1:* and the red one is slightly close to you
relative-p2, explicit, topological
47 is that right?
48 to my left from that red mug there is a yellow mug
relative-p1, explicit, topological
49 *P2:* hm...

Conversely, when looking for single objects that may be located anywhere on the entire table, for example, the speaker focuses on one object only that may be in a different part of the table than the one referred to in the previous utterance. There is no spatial continuum in the way the scene is processed and there may be several distracting objects that may lead to misunderstanding. Therefore, each description must be made more precise, both in the explicit definition of the FoR and through taking the perspective from which the reference is most identifiable. An example of this can be found towards the end of D1, before turn 115 (cf. Figure 4) where the participants decide to enumerate the mugs of each colour that they can see, P1 leads the enumeration and and describes the location of each object. However, the example also shows effects of continuity that is created by perceptual and discourse salience of objects, i.e. the way the scene is processed visually and the way it is described. In T117 "your left hand" is good landmark which attracts the FoR to P2 in the following spatial utterance in T119 but in T120 the FoR switches to P1 and in T121 back to P2. Turns T131-T136 show a similar object enumerating situation where FoR changes in every term and is also explicitly marked.

115 *P1:* my red ones are two in my first row (one of them close to katie)
relative-p1, explicit
116 *P2:*i mean there is a chance we both see a white that the other one is missing..
117 *P1:* one just next to your left hand
relative-p2, explicit
118 *P2:* yes
119 *P1:* and one on the third row from you slightly to your right
relative-p2, explicit
120 *P2:* is it directly behind the red mug on your left?
relative-p1, explicit
121 *P1:* no, much closer to you
relative-p2, topological
...
131 *P1:* and the blue ones are one on the second row from you, to the right from you
relative-p2, explicit
132 one slightly to my left
relative-p1, explicit
133 and one in front of katie in the first row
relative-katie, explicit
134 *P2:* yes, that's the same
135 *P1:* and the yellow are on between us to your far right
extrinsic
136 and one quite close to the corner on your left and katies right?
relative-p2, relative-katie, explicit

A switch between dialogue games tends to come with a switch of FoR. For example, in the following segment of D2, P1's FoR is selected initially to describe a row of cups closest to P1 and starting from their left to right (T14-T17). However, at T18 P1 initiates clarification. As P2 is information giver in this case the FoR is switched to theirs. Interestingly, the participants also switch the axis along which they enumerate objects (T21): starting at P2 and proceeding to P1, thus consistent from P2's perspective. At T26 a new clarification game is started and FoR changes to both P1 and P2, and at T32, after the participants exit both clarification games, P1 resumes the original game enumerating objects row-by-row and hence FoR is adjusted back to P1 accordingly.

14 *P1:* On my first row. I have from the left (your right): one red, handle turned to you but I can see it. A blue

cup next. Handle turned to my right. A white with handle turned to right. Then a red with handle turned to my left.
relative-p1, explicit
15 *P2:* first row = row nearest you?
relative-p1, explicit
16 *P1:* Yes.
17 *P2:* ok then i think we found a cup of yours that i can't see: the red with the handle to your left (the last one you mention)
relative-p1, explicit
18 *P1:* Okay, that would make sense. Maybe it is blocked by the other cups in front or something?
relative-p2
19 *P2:* yeh, i have a blue one and a white one, either of which could be blocking it
relative-p2
20 *P1:* Yes, I think I see those.
21 It looks almost like a diagonal line to me. From a red cup really close to you on your left, then a white, then the blue, then this missing red.
relative-p2
22 *P2:* blue with the handle to my left and white with the handle to my rigth/towards me a bit
relative-p2, explicit
...
26 *P1:* You know this white one you just mentioned. Is it a takeaway cup?
27 Because I think I know which cup that is but I don't see the handle.
28 *P2:* no, i was referring to the white handled cup to the right of the blue cup in the second row from you. its handle faces... south east from my perspective
relative-p1, relative-p2, explicit
29 the second row of cups from your end
relative-p1, explicit
...
32 *P1:* Shall we take my next row? Which is actually just a styrofoam cup. It's kinda marooned between the two rows.
relative-p1, explicit

(iii) Miscommunication and repair. We have already shown in the previous section that in line with (Mills and Healey, 2006), clarification triggers a change in FoR, with the explanation that clarification triggered a change of roles between the information giver and information receiver as well as introducing a different perceptual focus on the scene. However, during repair one would also expect that participants describe FoR explicitly more often. In the following example from D1, P1 is not sure about the location P2 is referring to. In T148 P2 explicitly describes the cup that can be found at that location using double specification of FoR. Information giver is thus providing more information that necessary to ensure precision of reference.

146 *P2:* so you see that yellow cup to be right on teh corner?
relative-p1
147 *P1:* Yes
148 A yellow cup, on my right your left, with the handle facing east to me, west to you.
relative-p1, relative-p2, explicit
149 *P2:* ok, from my perspective, there is at least a cup-sized gap between the edge of the table and the yellow cup
relative-p2, explicit
150 *P1:* Yes, I can say that too

As we have already seen, participants also use other strategies to reduce miscommunication, for example by enumerating objects that can be seen at any time of the conversation. From D1:

69 *P1:* so now I have 17 including the ones I've marked, how many do you have?
...
100 *P2:* so then again, it looks like we see everything we can
101 *P1:* yes, you still just got 17?
102 *P2:* yes

(iv) Explicit strategies Participants also devise strategies for processing the scene to find the missing objects. In (D1, T13) participants agree to use Katie's perspective as a reference. In (D2, T51 and following) they negotiate to split the table into a grid of 16 sub-areas where they label the columns with letters and rows with numbers. They negotiate the coordinates so that column labels A-D go from left to right and row numbers go from top to bottom relative to P2's view of the table. Hence, although they devise an extrinsic FoR with areas that they can refer to with coordinates they are forced to combine it with a FoR relative to P2 and therefore they create a more complicated system that involves two viewpoints. Interestingly, P1 clearly marked the axis labels on their printed sheet of the scene, which P2 did not, probably because the coordinate system was more difficult from P1's viewpoint. The negotiation of the coordinate system requires a lot of effort and involves referring to objects in the scene when negotiating where to start the lettering and numbering and how to place the lines for the grid. The participants finish the negotiation in T165, 114 turns later. However, although participants of D1 and D2 both negotiate on some reference perspective they do not use it exclusively as shown in Figure 4. One hypothesis that follows from these observations is that participants would use the reference (combining relative-katie and extrinsic) FoR in turns that involve greater information precision, that is those under repair as demonstrated in T119 of D2. Here the participants are negotiating where to draw the lines that would delimit different areas of the grid.

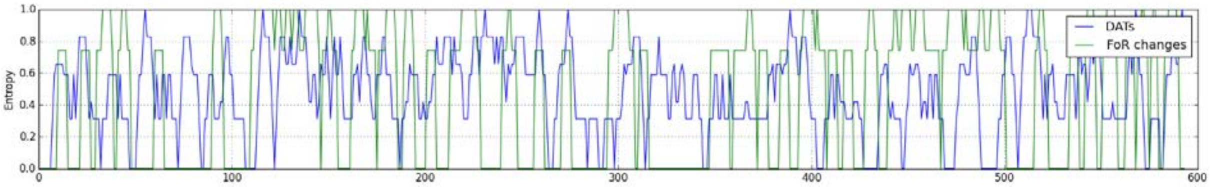


Figure 5: The entropy of DATs and FoR assignment calculated per each moving window of 5 turns. Both dialogues are combined into a single sequence and D2 starts in T158. Entropies were normalised by maximum observable entropy in the dataset.

105 *P2:* so, 2 could be in line with a can you see a blue cup, that is behind the A1 red cup?
relative-p1
...
110 *P1:* Yes. For me the blue cup is in front of the red cup. But yes.
relative-?, explicit
111 It has a handle that perhaps you can't see.
112 Since it is pointing south east for me.
relative-p1, explicit
113 *P2:* what do you mean by "in front of"
114 *P1:* Hmm
115 *P2:* closer to me or closer to you?
116 *P1:* Closer to you
relative-p1, explicit
117 *P2:* ok yep
118 *P1:* Okay
119 *P2:* i cna just see the handle almost pointing to A1
extrinsic

The excerpt shows that FoR itself may be open for repair. In T110 P1 corrects P2 in T105. P2's description contains FoR relative to P1, but P1 mistakenly takes a FoR relative to the landmark "the red cup" (i.e. intrinsic FoR). It is likely that this is because the red cup is very salient for P1 and allows P1 to project their orientation to the cup (the orientation of the FoR is not set by its handle). This is the only example where intrinsic FoR is used in the corpus and since it is repaired we do not count it as such. In T116 P1 comes to an agreement with P2.

3.5 FoR assignment over conversation

The preceding analysis of dialogue shows that FoR assignment is dependent on the type of communicative act or conversational game that participants are engaged in. The changes in perspective are dependent on factors that are involved in that particular game, for example the structure and other perceptual properties of the scene, the participants' focusing on the scene, their conversational role and availability of knowledge, the accommodated information so far, etc. To test whether the FoR assignment could be predicted only from the general dialogue structure we compared the entropy of the Dialogue Act tags with the entropy of the changes in FoR. As shown in Figure 5 there are subsections of the dialogue where the variability of DAs coincides with the variability of the FoR (i.e. where the entropy is high) but this is not a global pattern (Spearman's correlation rho = -0.36, $p = 0.383$). There are also no significant cross-correlations between the variables at different time lags. In conclusion, at least from our pilot data, we cannot predict the FoR from the general structure of conversational games at the level of DAs. This also means that there is no global alignment of FoR assignment and that this is shaped by individual perceptual and discourse factors that are part of the game.

4 Conclusions and future work

We have described data from a pilot study which shows how dialogue participants negotiate FoR over several turns and what strategies they use. The data support hypothesis (i) that there is no general preference of FoR in dialogue but rather this is related to the communicative acts of a particular dialogue game. Examining more dialogues would allow us to design an ontology of such games with their associated strategies which could be modelled computationally. Hypothesis (ii) that participants align over the entire dialogue, is not supported. Rather, we see evidence for local alignment. Hypothesis (iii) is also not supported: while misunderstanding may be associated with the use of different FoRs, there are also other dialogue games where this is the case, for example locating unconnected objects over the entire scene.

We are currently extending our corpus to more dialogues which will allow us more reliable quantitative analyses. In particular we are interested in considering additional perceptual and discourse features (rather than just DAs) to allow us to automatically identify dialogue games with particular assignments of FoR and therefore apply the model computationally.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly.

Holly Branigan, Martin Pickering, and Alexandra Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75:13–25.

Laura A. Carlson-Radvansky and Gordon D. Logan. 1997. The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37(3):411–437.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

Simon Dobnik, John D. Kelleher, and Christos Koniaris. 2014. Priming and alignment of frame of reference in situated conversation. In Verena Rieser and Philippe Muller, editors, *Proceedings of DialWatt - Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–52, Edinburgh, 1–3 September.

Nicholas D. Duran, Rick Dale, and Roger J. Kreuz. 2011. Listeners invest in an assumed other's perspective despite cognitive cost. *Cognition*, 121(1):22–40.

Eric N. Forsyth and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

Simon Garrod and Anne Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27:181–218.

Patrick G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg J. Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, MA, Aug.

Patrick G. T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PLoS ONE*, 9(6):e98598, June.

John D. Kelleher and Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

Boaz Keysar. 2007. Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, 4(1):71–84.

Willem J. M. Levelt. 1982. Cognitive styles in the use of spatial direction terms. In R. J. Jarvella and W. Klein, editors, *Speech, place, and action*, pages 251–268. John Wiley and Sons Ltd., Chichester, United Kingdom.

Stephen C. Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*. Cambridge University Press, Cambridge.

Xiaoou Li, Laura A. Carlson, Weimin Mou, Mark R. Williams, and Jared E. Miller. 2011. Describing spatial locations from perception and memory: The influence of intrinsic axes on reference object selection. *Journal of Memory and Language*, 65(2):222–236.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom, May.

Gregory Mills and Patrick G. T. Healey. 2006. Clarifying spatial descriptions: Local and global effects on semantic co-ordination. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, Potsdam, Germany, September.

Martin Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

Michael F. Schober. 1995. Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about locations? *Discourse Processes*, 20(2):219–247.

Holly A. Taylor and Barbara Tversky. 1996. Perspective in spatial descriptions. *Journal of Memory and Language*, 35(3):371–391.

Barbara Tversky. 1991. Spatial mental models. *The psychology of learning and motivation: Advances in research and theory*, 27:109–145.

Matthew E Watson, Martin J Pickering, and Holly P Branigan. 2004. Alignment of reference frames in dialogue. In *Proceedings of the 26th annual conference of the Cognitive Science Society*, pages 2353–2358. Lawrence Erlbaum Mahwah, NJ.

Learning non-cooperative dialogue policies to beat opponent models: “The good, the bad and the ugly”

Ioannis Efstathiou
Interaction Lab
Heriot-Watt University
ie24@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
o.lemon@hw.ac.uk

Abstract

Non-cooperative dialogue capabilities have been identified as important in a variety of application areas, including education, military operations, video games, police investigation and healthcare. In prior work, it was shown how agents can learn to use explicit manipulation moves in dialogue (e.g. “I really need wheat”) to manipulate adversaries in a simple trading game. The adversaries had a very simple opponent model. In this paper we implement a more complex opponent model for adversaries, we now model *all* trading dialogue moves as affecting the adversary’s opponent model, and we work in a more complex game setting: Catan. Here we show that (even in such a non-stationary environment) agents can learn to be legitimately persuasive (“the good”) or deceitful (“the bad”). We achieve up to 11% higher success rates than a reasonable hand-crafted trading dialogue strategy (“the ugly”). We also present a novel way of encoding the state space for Reinforcement Learning of trading dialogues that reduces the state-space size to 0.005% of the original, and so reduces training times dramatically.

1 Previous work

Recently it has been demonstrated that when given the ability to perform both cooperative and non-cooperative / manipulative dialogue moves, a dialogue agent can learn to bluff and to lie during trading dialogues so as to win games more often, under various conditions such as risking penalties for being caught in deception – against a variety of adversaries (Efstathiou and Lemon, 2014b; Efstathiou and Lemon, 2014a). Some of the adversaries (which are computer programs, not humans) could detect manipulation (with increasing probability as more manipulation moves occurred), but only had a simple opponent model which would try to estimate the preferences of the player agent. Furthermore, only specific moves (e.g. “I really need sheep”) affected the opponent model, and the setting was a simple 3-resource card-trading game. In this paper we model *all* trading dialogue moves as having effects on the adversary’s opponent model (i.e. “I will give you sheep for wheat” means that the adversary believes that the player needs wheat and doesn’t need sheep), and we work in the more complex setting of the Catan game (Afantenos et al., 2012).

2 Introduction

Work on automated conversational systems has been focused on cooperative dialogue, where a dialogue system’s core goal is to assist humans in their tasks such as buying airline tickets (Walker et al., 2001) or finding a restaurant (Young et al., 2010). However, non-cooperative dialogues, where an agent may act to satisfy its own goals rather than those of other participants, are also of practical and theoretical interest (Georgila and Traum, 2011), and the game-theoretic underpinnings of non-Gricean behaviour have been investigated (Asher and Lascarides, 2008). For example, it may be useful for a dialogue agent not to be fully cooperative when trying to gather information from a human, or when trying to persuade, argue, or debate, or when trying to sell something, or when trying to detect illegal activity, or in the area of believable characters in video games and educational simulations (Georgila and Traum, 2011; Shim and Arkin, 2013). Another arena in which non-cooperative dialogue behaviour is desirable is in negotiation (Traum, 2008), where hiding information (and even outright lying) can be advantageous. Dennett (Dennett, 1997) argues that a deception capability is required for higher-order intentionality in AI.

Machine learning methods have been used to automatically optimise *cooperative* dialogue management - i.e. the decision of what dialogue move to make next in a conversation, in order to maximise an agent's overall long-term expected utility, which is usually defined in terms of meeting a user's goals (Young et al., 2010; Rieser and Lemon, 2011). These approaches use Reinforcement Learning with reward functions that give positive feedback to the agent only when it meets the user's goals. This work has shown that robust and efficient dialogue management strategies can be learned, but until (Efstathiou and Lemon, 2014b), has only addressed the case of cooperative dialogue.

2.1 Corpus analysis

An example of the type of non-cooperative dialogue behaviour which we are generating in this work is given by our (dishonest) trading player agent A in the following dialogue:

A1: "I will give you a wheat and I need 2 clay"[A lies - it does not need clay but it needs wheat]
B1: "No"
A2: "I'll give you a rock and I need a clay"[A lies again and it actually needs rocks too, but it does not have any rocks to give]
B2: "No"
A3: "I'll give you a clay and I need a wheat
B3: "Yes"

Here, B is deceived into providing the wheat that A actually needs, because B believes that A needs clay (A asked for it twice) rather than wheat and rock (that it offered). Similar human behaviour can be observed in the Catan game corpus (Afantenos et al., 2012): a set of on-line trading dialogues between humans playing Settlers of Catan. We analysed a set of 32 logged and annotated games, which correspond to 2512 trading negotiation turns. We looked for explicit lies, of the form: *Player offers to give resource X (possibly for Y) but does not hold resource X* - such as in turn A2 in the above example.

11 turns out of 2512 were lies of this type. Since this corpus was not collected with expert players, we expect the number to be larger for more experienced negotiators. Other lies such as asking for a resource that is not really wanted, cannot be detected in the corpus, since the player's intention would need to be known.

2.2 Non-cooperative dialogues

Our trading dialogues are linguistically cooperative (according to the Cooperative Principle (Grice, 1975)) since their linguistic meaning is clear from both sides and successful information exchange occurs. Non-linguistically though they are non-cooperative, since they they aim for personal goals. Hence they violate Attardo's Perlocutionary Cooperative Principle (PCP) (Attardo, 1997). In the work below, the honest player agent proposes only sincere trades. It offers resources that are available and it asks for resources that it really needs. Hence it is learning to manipulate through legitimate persuasion (Dillard and Pfau, 2002; O'Keefe, 2002) and without any negative consequences. On the other hand, our dishonest player (see below) proposes false trades too, offers resources that are not available, and can ask for resources that it does not need. In other words, it can learn to manipulate based on lies and deception. We will show that both of the player agents can learn how to manipulate their adversaries through different but equally successful policies, by being cooperative on the locutionary level and non-cooperative on the perlocutionary level. In addition, we will present a hand-crafted naive agent who -like the honest player- is sincere, but does not learn how to use manipulation. In other words, it does not take into consideration at all the 'side effects' of its trading proposals, and we show that its performance is significantly lower than that of the two manipulative players.

2.3 Structure of the paper

We initially present the trading game "Catan" (section 3) and describe the version that we use for our experiments. All of the actions (trading proposals) that we use along with their manipulation mechanisms are presented and explained in detail. Section 4 presents the adversary and opponent model that we employ. We then propose a novel way of encoding (compressing) the state space for Reinforcement Learning (RL) with a tabular representation in Section 5, which reduces the training times dramatically. Then we present two Reinforcement Learning Agents (RLA) in Section 6 who -through honesty ("the good") and dishonesty ("the bad")- successfully learn how to use communicative manipulation (with every normal trading proposal). In Section 6.3 we investigate players without manipulation. Section 7 presents our experiments and detailed results are presented in Section 8.

3 The Trading Game "Catan"

To investigate non-cooperative dialogues in a controlled setting we used a 2-player version of the board game "Catan", which is a complex, sequential, non-zero-sum game with imperfect information. We call the 2 players the "adversary" and the "Reinforcement learning agent" (RLA). We also created a "hand-crafted agent" (HCA) for comparison. We assume that the adversary (see section 4) is affected by all the trading proposals of the learning agents, in such as way that it tries to stop the learning agents from getting the resources that they say they need. Intuitively, this is a basic aspect of adversarial behaviour.

The RLA or the HCA proposes trades to the adversary sequentially and tries to reach a goal number of resources (in the case of a city: 3 rocks and 2 wheat). There are four different goals that can be achieved in the normal "Catan" game: to build a road, a city, a settlement or buy a development card. Our RLA has also learned how to successfully trade in order to achieve all those goals but this paper is based on the example case of the city. There are five different resources to trade and the adversary only responds by either saying "Yes" or "No" to accept or reject the trade respectively. Currently we assume that the adversary has all of the resources available to give so it is up to the RLA or the HCA to use a successful strategy that will allow it to reach its goal. The learning agents start the game with a random number of resources (up to 7 of each resource) and therefore there are cases where the initial number of resources is insufficient to eventually reach their goal. The agents still learn how to get as close to the goal as possible (due to the reward function, see section 6).

3.1 Actions (Trading Proposals)

Trade occurs through trading proposals that may lead to acceptance or rejection from the adversary, and have deterministic and stochastic effects. We will first discuss the action's stochastic effect, that is whether or not the trade will be successful. In an agent's proposal (turn) only one 'give 1-for-1' or 'give 1-for-2' trading proposal may occur, or nothing (41 actions in total for the case of the dishonest RLA):

1. I will do nothing
2. I will give you a wheat and I need a timber
3. I will give you a wheat and I need a rock

...

40. I will give you a brick and I need two rocks
41. I will give you a brick and I need two sheep

In contrast to the case of the dishonest RLA, the cases of the honest RLA and the naive HCA consist of 17 of the above actions because they ask only for goal resources (rock and wheat). The adversary responds by either saying "Yes" or "No" to accept or reject the learning agent's proposals. Each of these actions affects the adversary's opponent model as described below.

3.2 Manipulation through trading actions

We assume that all of the above trading proposals (apart from "I will do nothing") affect the opponent model of the adversary. Hence a trading proposal may or may not lead to a trade (the action's stochastic effect) as we saw, but it will definitely affect (action's deterministic effect) the adversary's belief model. Here we will discuss each action's deterministic effect. Each of the trading proposals consists of two parts: the offered resource and the wanted one(s). The adversary's opponent model is affected by both of these parts – for example the more often the agent insists on asking for wheat, the less the adversary will be eager to give it. Hence the agents need to learn how to appropriately use this effect in order to successfully manipulate the adversary and reach the goal number of resources.

4 The Adversary and its Opponent model

The adversary remains the same in all of our experiments. However other adversary and opponent models are clearly possible. We created this as a simple implementation of the intuition that a rational adversary will act so as to hinder other players in respect of their expressed preferences.

Opponent models (OM) with hindering abilities have previously been shown to be important in games such as the "Machiavelli" card game

(Bergsma, 2005). Hence our adversary is using an opponent model that is based on hindering the LA's preferences, as the LA expresses its preferences through trading proposals and this is the only information that the adversary receives. Since opponent modeling is focused on using knowledge about other agents to improve performance, the adversary therefore hinders the LA's announced preferences (trading proposals).

Our model is inspired by this approach to OM and uses knowledge (from the LA's announcements) in an effort to improve its performance. Unlike the OM (Carmel and Markovitch, 1993; Iida et al., 1993a; Iida et al., 1993b) or the PrOM search model of (Donkers et al., 2001) though, it does not explicitly predict the moves of the LA, but the history of those moves are used to direct the adversary's future responses.

The adversary therefore uses an opponent model which directs its responses to the other agent's (RLA or HCA) trading proposals. Every time that an agent utters a trading proposal, probabilities of the adversary giving resource types change accordingly (details below), and therefore the adversary becomes more or less eager to give some resources than others. It does this because it tries to hinder the other players from acquiring the resources that they ask for. For instance, if an agent insists on asking only for wheat then the probability that it will be given becomes very low (the adversary considers it now as valuable), but the relative probability that it will get one of the other four resources increases.

However, the adversary also takes into consideration what the agent offers to *give*, so the more an agent keeps offering a resource the more likely becomes for the adversary to give it too (it considers the resource as less valuable).

In detail, at the beginning of each trading phase the probabilities that represent the adversary's willingness to give each of the resource types start at 50%. When the agent asks for a resource then the probability to give that particular resource is reduced by either 8% or 12% (if it is a 'give 1-for-1' or 'give 1-for-2' trade proposal respectively), and the probability of giving the four other resource types increases accordingly. The probability of giving the offered resource also increases by 8%. We experimented with a variety of different increments, and very similar results were obtained to those presented below, so there is nothing particularly hinges on the 8% figure.

Due to this opponent model, it is possible to manipulate the adversary into eventually giving resources that are needed, if the right trading proposals are made.

5 The State Encoding Mechanism

To overcome issues related to long training times and high memory demands, we implemented a state encoding mechanism that automatically converts all of our trading game states to a significantly smaller number states in a compressed representation. The new state representation takes into consideration the distance from goal and the availability of the resource, as well as its quality (goal or non-goal resource) and uses 7 different characters. The agent's state consists of the numbers of the five resources that it currently has available. In the case of the city, it needs wheat and rocks. That means two out of five resources are goal resources and therefore they can be represented by 'G' (goal) when their number is equal to the goal amount, 'N' (null) when their number is 0, 'M' (more) when their number is more than the goal-quantity, and '1' or '2' when the distance from the goal quantity is 1 or 2 respectively. The 3 non-goal resources are represented by 'Z' (zero) when they are 0 and 'A' (available) when they are more than 0.

For example, the state $\langle 1, 4, 3, 0, 2 \rangle$ would be encoded to $\langle 1, A, G, Z, A \rangle$. The numeric state space of our problem has 8 x 8 x 8 x 8 x 8 (=32,768) states that are encoded to only 4 x 2 x 5 x 2 x 2 (=160) states. This is reduced to 0.005% of the original size of the state space. With this method and despite the fact that the representation still remains tabular, in all of our experiments 3 million training games required only around 10 minutes to finalize. The performances were very successful too as the logic is still based on the precision of the RL tabular representation.

6 The Reinforcement Learning Agents (RLA)

As we discussed earlier the game state is represented by the RLA's encoded set of resources (see section 5). The RLA plays the game and learns while perceiving only its own set of resources. It is aware of its winning condition in as much as it experiences a large final reward when reaching this state. It learns how to achieve the goal state through trial-and-error exploration while playing repeated games. Each game consists of up to 7 trading proposals, but nothing particularly hinges upon this number – we have experimented with a number of different length constraints, and obtained similar results. The agent is modelled as a Markov Decision Process (Sutton and Barto, 1998): it observes states, selects actions according to a policy, transitions to a new state (due to the adversary's response), and receives rewards at the end of each game. This reward is then used to update the policy followed by the agent using the SARSA(λ) algorithm.

As we see in Figure 1, it learns to win 96.8% of the time (not 100% due to the cases with insufficient initial resources).

6.1 Reward function

The reward function used in all the experiments takes into consideration the number of trading proposals made and the distance from the goal, as well as trading success. In detail, the reward function that is used is: + 10,000 (if trading successful) $-(1,000$* proposals) $-(1,000$* distance).

6.2 Training parameters

The agents were trained using a custom SARSA(λ) learning method (Sutton and Barto, 1998) with an initial exploration rate of 0.2, which gradually decays to 0, and a learning rate α of 1, which also gradually decays to 0 by the end of the training phase. After experimenting with the learning parameters we found that with λ equal to 0.9 and γ equal to 0.9 we obtain the best results for our problem and therefore these values have been used in all of the experiments that follow.

6.3 Initial cases with no manipulation / cooperative adversary

Before we examine the cases with manipulation and the adversary's opponent model, we first explore the case of learning a trading policy for adversaries that do not have an opponent model and thus do not try to hinder the learning agent. This adversary always accepts an agent's trading proposal, and so this serves as an initial proof-of-concept of the extent to which the game is winnable by the learning agents if the adversary is being fully cooperative.

Here the RLA learned how to successfully trade in the full version of the "Catan" game for every goal case. These include building a road, a city, a settlement, or a development card. The different goals are different numbers and types of resources that the RLA needs to gather in order to win.

The RLA has located a successful policy for each one of those cases, showing that the cooperative version of the game is solvable as an MDP problem. It has identified and taken advantage of the power of the 'give 1-for-2' over the 'give 1-for-1' trades and therefore it uses them much more frequently (with a ratio of around 75% over 25% for the 'give 1-for-1'). The adversary that it plays against does not have an opponent model, the learning agent's trading proposals do not affect it, and the adversary always accepts them. Hence we initially show that RL is capable of successfully learning how to trade in this version of the game (with every different goal) while learning to also exploit the 'give 1-for-2' trading proposals.

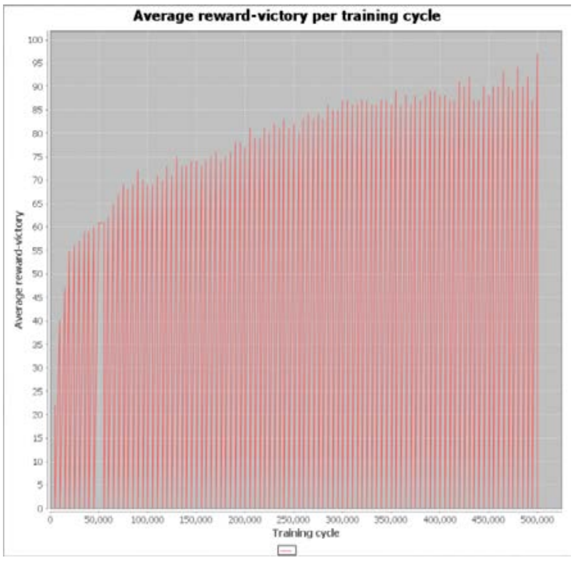


Figure 1: *Learning Agent's reward-victory graph in 500 thousand training games of Initial Experiment: building a city, cooperative adversary.*

6.4 The Honest Reinforcement Learning Agent - "The Good"

The honest RLA only asks for resources that it really needs (therefore it is restricted to 17 out of the 41 actions). It is a sincere RLA and it only proposes a trade after it has checked that the offered resource is indeed available. However, the fact that it still learns how to successfully manipulate (legitimately persuade) the adversary under those honest constraints, and in a continuous non-stationary MDP environment due to the ever-changing adversarial belief model (i.e. the environment's dynamics can change after an action is selected), makes the outcome surprising. In the experiments that follow we will see that it locates a honest way of persuading its adversary.

6.5 The Dishonest Reinforcement Learning Agent - "The Bad"

The dishonest RLA can ask for resources that it does not need (therefore it uses all of the 41 actions). It can also propose trades without checking if the offered resource is available. If such a deceitful trading proposal gets accepted by the adversary, the RLA then refuses to actually make the trade. Thus its learning process is a harder Reinforcement Learning task than that of the honest RLA (since it has more actions). However, it still learns how to successfully manipulate (deceive) the adversary under those dishonest conditions, and in a continuous non-stationary MDP environment due to the ever-changing adversarial opponent model as above, resulting on a surprisingly equal performance with that of the honest RLA. As we will see in the experiments that follow, its strategy is based on the use of lies.

6.6 The Naive Hand-Crafted Learning Agent - "The Ugly"

This agent is not a learning agent but instead uses a hand-crafted naive strategy. In detail, it uses a reasonable way of proposing trades by checking the availability of the resources that it does not need and offers them for those that it needs in an equi-probable manner. The reason that we call it naive (as well as "ugly") is because it does not take into consideration the fact that its trading proposals affect the adversary's opponent model and -instead of learning that- it just keeps following the same naive rule-based strategy. This agent is a baseline case and despite the fact that its strategy is quite sensible, we show that it is significantly worse than that of the two manipulative RLAs.

7 Experiments

All agents are compared in respect of their win rates, which is the percentage of trading games in which they achieve their goal (in this case, to get the resources required to build a city). The y-axes of the graphs below represent this quantity (which we also refer to as "success rate" or "reward-victory").

7.1 Naive HCA vs. Adversary: Experiment 1 (Baseline)

The naive HCA played 3 million games against the Adversary in Experiment 1. This is our baseline case for comparison. The agent's trading proposals affect the opponent model of the adversary but the agent is unaware of that and therefore it does nothing about it. It just keeps playing the game based on the naive but reasonable strategy discussed in Section 6.6.

7.2 Honest RLA vs. Adversary: Experiment 2

In this experiment we trained the honest RLA against the adversary in 3 million games. The RLA's trading proposals affect the opponent model of the adversary and we show that, despite the honest constraints, the honest RLA can learn how to successfully manipulate the adversary. Ultimately we show that the performance is better than that of the baseline case in Experiment 1. The performance of the Honest RLA before training (i.e. random action selection) is about 21%.

7.3 Dishonest RLA vs. Adversary: Experiment 3

In this experiment we trained the dishonest RLA against the adversary in 3 million games. The RLA's trading proposals again affect the opponent model of the adversary and we show that the dishonest RLA can learn how to successfully manipulate it. As above, we show that the performance is better than that of the baseline case in Experiment 1. Furthermore, we explore how well this deceitful RLA performs compared to the previous honest one, who legitimately persuades. The performance of the Dishonest RLA before training (i.e. random action selection) is about 4%.

8 Results

The RLAs were trained on 3 million games against the Adversary. Their policies were then tested in 20,000 games. The HCA played 3 million games too against the same adversary. As there was no learning, no testing games were played because its performance remained stable throughout the 3 million games as we will see below.

8.1 Naive HCA: Experiment 1

The naive HCA has a win rate of only 25.3%. Its strategy focuses on 50% of the time asking for wheat by offering each one of its available unwanted resources in turn, or 50% of the time asking for rocks using the same technique.

8.2 Honest RLA: Experiment 2

The honest RLA scored a winning performance of 35.8%, see Figure 2, starting from 21.1% (which is the performance of random action selection). Its strategy focuses on asking initially for either wheat, until it gathers rocks, or for rocks until it gathers wheat that needs to build a city (2 wheat and 3 rocks are required). It also mainly offers resources that it needs (goal ones) -and has available- instead of non-goal ones as it will become then easier to get them back. This honest persuasive strategy proved to be very effective against the adversarial hindering policy.

8.3 Dishonest RLA: Experiment 3

The dishonest RLA scored a winning performance of 36.2% after 3-million training games, and may improve with further training (see Figure 3), starting from only 4.2%. That clearly shows that its task was much harder than that of the honest RLA in Experiment 2, who started from 21.1%, as it has to understand how to effectively manipulate through all of the 41 actions (rather than the 17 honest actions which ask for goal resources only). Nevertheless its very effective learned strategy mainly focuses on the use of lies. It asks especially for resources that it does not need only for the sake of manipulation (deception) and it offers resources that it does not have for the same purpose. The type of the offered resources in this case are mainly goal ones again (as above) and the fact that this RLA can lie about their availability makes such offers even more frequent than before. This dishonest strategy proved to be equally effective with that of the honest RLA though.

Both of the RLAs (as we saw in Experiment 2 too) managed to learn successful strategies despite the fact that there are cases where the initial resources are insufficient to reach the goal within 7 proposals. They both realized again (as in our Initial Experiment, section 6.3) the power of the 'give 1-for-2' over the 'give 1 for-1' trades and they used them more often. Hence, in some cases they manage to approach their goals even with insufficient initial resources. By comparing the two manipulative cases to that of Experiment 1 we show that manipulation (through legitimate persuasion [Experiment 2] or deception [Experiment 3]) can be successfully learned by our RLAs and outperform by 11% a naive but reasonable strategy.

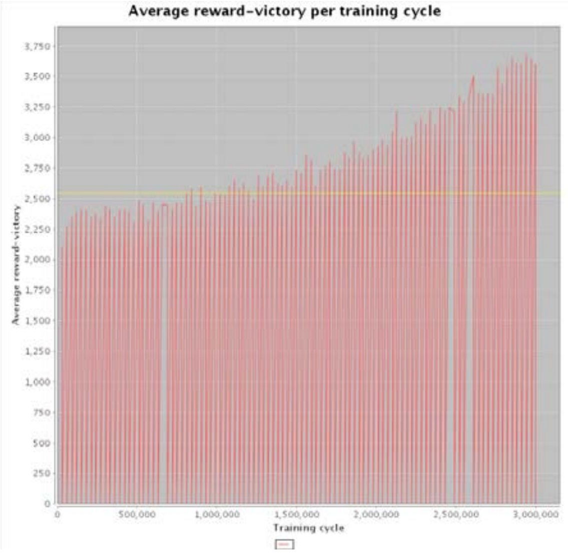


Figure 2: *Honest RLA's reward-victory graph in 3 million training games (experiment 2). Yellow horizontal line = Baseline performance.*

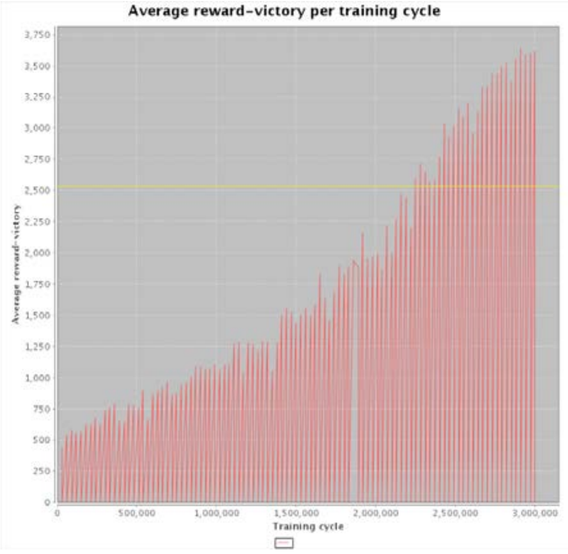


Figure 3: *Dishonest RLA's reward-victory graph in 3 million training games (experiment 3). Yellow horizontal line = Baseline performance.*

9 Discussion: a Non-Stationary MDP problem

Our Experiments 2 and 3 also show that RL is capable of learning successful policies even in the case where the environment's dynamics change (maximum of 7 times per game) and each action (trading proposal) has a stochastic effect (that of

Exp.	Learning Agent policy	Adversary policy	Agent's wins
Initial	SARSA + Honest actions	Accepts every trade	96.8%
	Random Honest actions	Hinders agent's preferences	21%
	Random Dishonest actions	Hinders agent's preferences	4%
1	Hand-Crafted Naive Honest (Baseline)	Hinders agent's preferences	25.3%
2	SARSA + Honest actions	Hinders agent's preferences	35.8%*
3	SARSA + Dishonest actions	Hinders agent's preferences	36.2%*

Table 1: *Performance (% wins) in 20 K testing games, after training. (*= significant improvement over baseline,* $p < 0.05$*)*

a possible trade) and a deterministic effect (that on adversary's opponent model). Every time the honest or dishonest RLA proposes a trade, the opponent model of the adversary changes as we have seen. That means the environment changes too (as the adversary is a part of it according to the RLA's perspective) and therefore makes our problem a non-stationary MDP (da Silva et al., 2006). Despite the fact that only the RLA's actions are responsible for those changes and so the problem may be solved by recasting it into a stationary one through state augmentation (Choi et al., 2001), our case is more complex. This is because our RLA's actions affect the environment in two different ways (through their stochastic and deterministic effects). Furthermore, the environment (adversary) responds to trading proposals based on the history of the deterministic effects of the actions (trading proposals' effect on adversary's belief) up to that point. In other words, the same action (trade) may have different effects due to the deterministic effects on the environment (changes of the adversary's opponent model) of the actions that preceded it. There are successful combinations between these two different kinds of effects that the RLA has managed to identify and learn how to effectively use, originating from the multi-dimensions (manipulative dimensions) of the problem. It is therefore an interesting multi-dimensional non-stationary MDP case that we have shown to be solvable by RL, which suggests that trading proposals in dialogue evoke non-stationary beliefs in our everyday negotiations. We demonstrated that phenomenon with the realistic assumption that the adversary's opponent model is affected by all normal trading actions.

10 Discussion: Discourse Studies

Our results also bring an important argument of Van Dijk (van Dijk, 2006) to light, according to which there is an everyday conventional inference of dishonesty from manipulative acts. That negative effect cannot be taken for granted though as manipulation according to Dillard and Pfau, as well as O'Keefe (Dillard and Pfau, 2002; O'Keefe, 2002) also occurs through legitimate persuasion. This is what our RL work suggests too. Hence we emphasize the significance of Attardo's perlocutionary cooperation as before.

11 Conclusion & Future Work

In this paper we implemented an opponent model for adaptive adversaries, and modelled *all* trading dialogue moves as affecting the adversary's opponent model. We worked in the complex game setting of Catan and we showed that agents can learn to be legitimately persuasive ("the good") or deceitful ("the bad"). We achieve up to 11% higher success-rates than a reasonable hand-crafted trading dialogue strategy ("the ugly").

We also presented a novel way of encoding the state space for Reinforcement Learning of trading dialogues that reduces the state-space size to 0.005% of the original, and so reduces training times dramatically.

In future work we will further investigate complex non-cooperative situations, and evaluate the performance of such learned policies in games with humans, by integrating this work with jSettlers (Thomas and Hammond, 2002).

Acknowledgements

This work is partially funded by the European Research Council, grant no. 269427 (STAC project). Thanks to Heriberto Cuayáhuitl for help with the corpus analysis.

References

S Afantenos, N. Asher, F. Benamara, A. Cadilhac, C Dégremont, P Denis, M Guhe, S Keizer, A. Lascarides, O Lemon, P. Muller, S. Paul, V. Popescu, V. Rieser, and L. Vieu. 2012. Modelling strategic conversation: model, annotation design and corpus. In *Proc. 16th Workshop on the Semantics and Pragmatics of Dialogue (Seinedial)*.

N. Asher and A. Lascarides. 2008. Commitments, beliefs and intentions in dialogue. In *Proc. of SemDial*, pages 35–42.

S. Attardo. 1997. Locutionary and perlocutionary cooperation: The perlocutionary cooperative principle. *Journal of Pragmatics*, 27(6):753–779.

M.H.J. Bergsma. 2005. Opponent Modeling in Machiavelli. B.s. thesis, Maastricht University, the Netherlands.

D. Carmel and S. Markovitch. 1993. Learning models of opponent's strategies in game playing. In *Proceedings AAAI Fall Symposium on Games: Planning and Learning*, pages 140–147. The AAAI Press.

Samuel P.M. Choi, Dit-Yan Yeung, and Nevin L. Zhang, 2001. *Sequence Learning - Paradigms, Algorithms, and Applications*, chapter Hidden-Mode Markov Decision Processes for Nonstationary Sequential Decision Making. Springer-Verlag.

Bruno C. da Silva, Eduardo W. Basso, Ana L.C. Bazzan, and Paulo M. Engel. 2006. Dealing with Non-Stationary Environments using Context Detection. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*.

Daniel Dennett. 1997. When Hal Kills, Who's to Blame? Computer Ethics. In *Hal's Legacy:2001's Computer as Dream and Reality*.

James Price Dillard and Michael Pfau. 2002. *The Persuasion Handbook: Developments in Theory and Practice*. SAGE Publications, Inc.

H. H. L. M. Donkers, H. J. Van Den Herik, and J. W. H. M. Uiterwijk. 2001. Probabilistic opponent-model search. *Information Sciences*, 135:123–149.

Ioannis Efstathiou and Oliver Lemon. 2014a. Learning to manage risk in non-cooperative dialogues. In *Proc. SEMDIAL*.

Ioannis Efstathiou and Oliver Lemon. 2014b. Learning non-cooperative dialogue behaviours. In *SIGDIAL*.

Kallirroi Georgila and David Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Proc. INTERSPEECH*.

Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3.

H. Iida, J.W.H.M. Uiterwijk, H.J. van den Herik, and I.S. Herschberg. 1993a. Opponent-model search. Technical report cs 93-03, Universiteit Maastricht.

H. Iida, J.W.H.M. Uiterwijk, H.J. van den Herik, and I.S. Herschberg. 1993b. Potential applications of opponent-model search. part 1: The domain of applicability. *ICCA Journal*, 16(4):201–208.

Daniel O'Keefe. 2002. *Persuasion: Theory and research (2nd Edition)*. SAGE Publications, Inc.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement Learning for Adaptive Dialogue Systems: A Data-driven Methodology for Dialogue Management and Natural Language Generation*. Theory and Applications of Natural Language Processing. Springer.

J. Shim and R.C. Arkin. 2013. A Taxonomy of Robot Deception and its Benefits in HRI. In *Proc. IEEE Systems, Man, and Cybernetics Conference*.

R. Sutton and A. Barto. 1998. *Reinforcement Learning*. MIT Press.

R. Thomas and K. Hammond. 2002. Java settlers: a research environment for studying multi-agent negotiation. In *Proc. of IUI '02*, pages 240–240.

David Traum. 2008. Extended abstract: Computational models of non-cooperative dialogue. In *Proc. of SIGdial Workshop on Discourse and Dialogue*.

Teun A. van Dijk. 2006. Discourse and manipulation. *Discourse & Society*, 17(2):359–383.

M. Walker, R. Passonneau, and J. Boland. 2001. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Steve Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. 2010. The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174.

Exploring age-related conversational interaction

Jeroen Geertzen
Dept. of Theoretical & Applied Linguistics
University of Cambridge
`jg532@cam.ac.uk`

Abstract

Does the way we interact in conversation changes as we get older? This paper presents a corpus-based study investigating this question by looking into age-related differences in spontaneous spoken dialogue. Conversations from the Switchboard corpus were analysed using n-gram models, relating the communicative actions as encoded by dialogue acts to speaker age. Results show that older interlocutors generally address the conversation itself (rather than other topics) less than younger adults. Age differences are also reported on feedback strategies, most notably that even though younger interlocutors produce more backchannels overall, older interlocutors use backchannels more often when taking turns.

1 Introduction

Numerous studies on spoken language production have documented language production changes across the life span (see (Mortensen et al., 2006) for an overview). Much of this research is highly controlled in nature and has focused on word level production (see (Burke and Shafto, 2004) for a review). Yet, relatively little research has addressed the question whether normal aging affects the way interlocutors behave and interact with each other in conversation. This, in itself, is not surprising as conversation is difficult to investigate in a controlled manner unless it involves very specific tasks that often make the experimental setting rather artificial.

There is some evidence, however, that suggests that age may have an effect. For instance, (Hupet et al., 1993) showed, by repeatedly asking pairs of interlocutors to discuss how to arrange complex figures in a particular order, that older adults are less likely to take previously shared information into account than younger adults. But many aspects that could be affected by age differences have not been looked at, leaving many questions unanswered.

Do older interlocutors spend more or less effort to address the communication process itself by providing communicative feedback as compared to addressing the topics or tasks under discussion? If there is an aging effect present, what forms of communicative feedback feature such effects? For instance, the use of 'backchannels', short verbal expressions such as 'uh-huh' or 'yeah' are considered to provide implicit positive feedback, whereas expressions such as 'are you listening?' and 'what?' signal communicative problems (Allwood et al., 1992; Bunt, 1994). Age related differences in how and how often feedback is produced could reveal differences in the communicative performance or strategies involved.

The aim of this paper is to investigate whether, and if so how, age affects the way interlocutors interact with each other in spontaneous spoken dialogue by analysing interaction patterns using recorded and annotated telephone conversations from the Switchboard I corpus (Godfrey et al., 1992).

Section 2 provides further background by elaborating on key aspects of spoken dialogue, while Section 3 reports on the Switchboard data, its speaker characteristics, and the annotations used. Section 4 provides the results obtained, which are discussed in Section 5. Conclusions and future work are described in Section 6.

2 Background

The communicative functions of utterances that speakers contribute to a dialogue are usually characterized by dialogue acts, which can be taken to characterise the interaction and constitute sequences such as Question-Answer or Greeting-

Greeting, or more complex patterns. For instance, interlocutor B could respond to a yes-no question posed by interlocutor A by asking A a clarification question and receiving A's reply before finally answering the yes-no question.

Besides communicating about some underlying task or activity that drives the dialogue, interlocutors have to attend to various aspects of the communicative process itself as to keep the communication going in a sufficiently smooth way. This process, often referred to as *dialogue management*, may involve signalling changing time constraints (time management), assigning sender and receiver roles (turn management), or monitoring mutual attention and understanding. Such engagement gives rise to communicative activities such as taking the turn, showing attention, signalling a misunderstanding, and establishing joint attention on the conversational topic under focus. Clark (1996) describes this distinction by means of two communicative tracks: communicative acts in the first track are concerned with the presentation of the information, whereas meta-communicative acts in the second track are concerned with the communication itself. One important aspect is, therefore, how much of the contributions that interlocutors make involve dialogue management, and little dialogue management is arguably an indicator for well going conversation.

At the same time, a conversation without any communication management is not necessarily perceived as more fluent and effective than a conversation with little communication management, which motivates the distinction between communication management that helps to keep a non-problematic interaction continue sufficiently smoothly and dialogue management that is typically used when problems in the communication arise. *Backchannels* are typical of the former category and are short utterances, such as 'uhu' and 'hm-mm', produced by listeners to signal understanding and indicating that they are paying attention, encouraging speakers to continue (Duncan, 1972). They are considered to make the conversation go smoothly. The latter category is mostly represented by *explicit negative feedback*, which concerns utterances conveying meta-communicative acts that indicate a problem in the understanding that may occur for various reasons such as problems in perception (e.g. "I cannot hear you") or interpretation (e.g. "What did you mean?").

For each of the key communicative aspects discussed, there is no clear theory-driven hypothesis that has been proposed. It is generally assumed that healthy aging in (late) adulthood tends to be accompanied by a subtle decline in cognitive and perceptual functioning. Where this affects spoken dialogue, a recent study has established that older adults show more difficulty in following conversation, largely accounted for by perceptual functioning (Murphy et al., 2006). This may cause older adults to engage more often in dialogue management and produce relatively more signals of non-understanding. At the same time, research in human-computer interaction has found that older adults use a wider range of speech acts allowing more flexible interaction patterns (Georgila et al., 2008).

The question whether there are age-related differences in interaction, and the lack of strong hypotheses makes this an exploratory study in which not only the small set of dialogue management functions will be included, but also other less important aspects are considered.

3 Material and method

3.1 Speech data

The speech data for this study come from the Switchboard I Corpus (Godfrey et al., 1992), a collection of about 2,400 dialogues among 543 speakers from all areas of the United States on a wide range of preferred conversational topics. Speakers sharing an interest in the same topic were paired such that no two speakers would converse together more than once, and no one spoke more than once on a given topic. Basic demographic information was collected for each speaker, including age.

A part of the Switchboard corpus was manually annotated with a set of dialogue acts (Jurafsky et al., 1997) and released as the Switchboard Dialog Act Corpus. It covers 1,155 five-minute dialogues comprising over 200K utterances and 1.4 million words. The dialogue act tagset that was used, SWBD-DAMSL, is based on the Discourse Annotation and Markup System of Labeling (DAMSL; (Allen and Core, 1997); (Core and Allen, 1997)) and contains 220 tags that are clustered into 42 larger classes. A dialogue excerpt with labelled utterances is illustrated in Table 1.

Table 1: Sample from dialogue `sw_0811_2278`

	Utterance	*Dialogue act*
B	Okay,	Other (`o`)
B	um, so, um, do you have any favorite teams?	Yes-No-Question (`qy`)
A	Well, I kind of like them all.	Affirmative non-yes answers (`na`)
A	I played for about eighteen years, all the way through college, and then, uh, kind of hung them up after college,	Statement-non-opinion (`sd`)
A	but, $<$ *laughter* $>$	Abandoned or Turn-Exit (`%`)
B	Oh, I was going to say, you played pro ball,	Statement-non-opinion (`sd`)
B	right?	Tag-Question (`^g`)

3.2 Speakers

The sample for this study consists of speech from 438 different speakers, taken from the 1,155 unique conversations in the corpus. The speakers are between 20 and 68 years old (M=37.6; SD=10.9) and are balanced in gender. Speakers could take part in multiple dialogues: 31% participated only in a single dialogue; 28% participated in two to five dialogues, and 41% participated in more than five dialogues.

3.3 Analysis

The first part of the analysis is based on measuring production rates of individual dialogue acts that target key aspects of the interaction.

In the SWBD-DAMSL and DAMSL annotation scheme, the Information Level layer indicates whether a contribution to the dialogue is about the task, about the management of the task (`^t`), or about the communication (`^c`);

As for back-channels, we focus on two specific tags in DAMSL targetting this phenomenon: general backchannels (`b`) and backchannels in question form (`bh`), such as really? or yeah?. Explicit negative feedback is indexed by “Signal-non-understanding” which in DAMSL are explicitly marked by (`br`) and (`br^m`) tags.

For each of the 438 speakers, production rates were computed for each dialogue act tag by dividing the frequency of the dialogue act tag by the sum of frequencies of all dialogue acts. As the five-minute fragments are rather short, production rates were not first averaged for each dialogue but computed based on all dialogue contributions by a particular speaker. Presence of age effects were then tested by correlating production rates of dialogue acts with interlocutor’s age (in years). Pearson correlation coefficients (r) and corresponding p-values (p) are reported.

Correlation analysis has the advantage that the whole age range is modelled as opposed to binning interlocutors in an ‘old’ and ‘young’ group according to a more or less arbitrary ranges.

With the same approach as for individual dialogue acts, also interaction patterns of subsequent dialogue acts were analysed to get a better picture of the conversational interaction an interlocutor is involved in. To include turn boundaries as well, turn changes were marked explicitly as turn-beginnings (denoted by $|B|$) and turn-endings (denoted by $|E|$) with respect to a specific interlocutor, e.g. `qy+`$|B|$`+na+sd+%+`$|E|$`+sd+^g` for speaker B. To describe subsequences of dialogue acts, n-gram language models of various orders were estimated: bigrams, trigrams, and 4-grams.[1]

With the individual dialogue acts, we have no clear hypotheses to test. More problematically, the risk of Type I errors (false positives) in significance testing increases with the number of hypotheses being addressed, which are numerous when looking for prominent dialogue act n-grams, making the significance level increasingly meaningless. Multiple comparisons can be taken into account by various adjustments, such as dividing the usual alpha of 0.05 by the number of hypotheses involved (Bonferroni, 1935), but sound hypothesis identification as well as (replication) testing can be achieved by data splitting (Dahl et al., 2008). This technique involves randomly splitting the data into two parts: one for hypothesis formulation and one for hypothesis testing. Only hypotheses that are identified by a p-value below

---

[1] For instance, a bigram language model would involve pairs {`o+qy, qy+na, a+sd, sd+%, %+sd, sd+^g`} and is expected to pick up on adjacency pairs.

alpha-level in the first part which then are also tested as significant in the second part are considered to be truly significant as well as replicable, and be reported as estimated over all data.

4 Results

4.1 Individual dialogue acts

It may be argued that the more an interlocutor is engaged in communication management, the less fluent the dialogue tends to become, and age related differences may emerge at Information-level. The correlations of age with the relative frequencies of the labels in Information-level are listed in Table 2, and show that older speakers are generally less involved in communication management.

Table 2: Significant correlations with Information-level: communication (C) and task (T), with task management not replicable

	$r(436)$	p	p_1	p_2
C	-0.199	0.00003	0.0022	0.0045
T	0.186	0.00008	0.0046	0.0067

Table 3: Correlation of age with backchannels (B), which in question form are not replicable

	$r(436)$	p	p_1	p_2
B	-0.187	0.00008	0.0031	0.0073

Pearsons correlations (Table 3) show that relative to all produced dialogue acts, the use of backchannels decreases with age. This is one of the stronger correlation that was found, and is relevant considering that backchannels account for around 19% of all dialogue acts produced (and backchannels in question form for around 1%).

Generally, older interlocutors produced fewer questions and more statements. They tend to signal less non-understanding, even though this turned out not to be replicable (Table 4). Also, older speakers tend to use fewer hedges: expressions that intend to diminish the confidence or certainty of a statement or answers that the speaker made, such as ‘I guess’ or ‘If I am not mistaken’.

4.2 Dialogue act n-grams

Dialogue act sequences involving acts from both interlocutors were described by means of n-grams of which the production rate of a n-gram was calculated as its frequency divided by the cumulative frequency of n-grams of the same length. Production rates were correlated with age, and a selection of the strongest correlating n-grams are listed in Table 5.

Table 4: Correlation with Questions (Q), Hedges (H), Statements-non-opinion (S), and signal-non-understanding (SNU), which cannot be reproduced

	$r(436)$	p	p_1	p_2
Q	-0.173	0.00026	0.0102	0.0078
H	-0.169	0.00038	0.0095	0.0139
S	0.136	0.00436	0.0451	0.0400
SNU	-0.116	0.01539	0.0438	0.2343

Older speakers produce fewer acknowledgements in the form of backchannels following a statement of the dialogue partner (5.1), while receiving more elaborate backchannels or continuers that express appreciation, such as ‘That sounds great.’ or ‘I can imagine’ (5.2). Also generally, older speakers produce fewer backchannels and younger speakers produce more (5.7). Furthermore, age shows a negative correlation with hedges followed by an opinion statement (5.6).

Older speakers produce more tag questions, such as ‘So you like music, don’t you?’, which are then followed by a descriptive or narrative statement which acts as a negative answer, such as ‘I do not’, produced by the other interlocutor (5.3).

5 Discussion

As results suggest, younger adults produce more backchannels in various circumstances, indicating explicitly the active monitoring of the partner’s production. This finding is in line with existing work. For instance, (Kemper et al., 1998) report a significant age effect in a referential communication task (a map task as in (Anderson et al., 1991)) in which young adults instructed older adults to reproduce a map. Also in other tasks, such as describing to each other a mutually experienced event such as holidays, younger adults produced more backchanneling than older adults (Gould and Dixon, 1993). This has been explained as an increased “willingness and ability to take on the cognitively demanding task of dividing one’s attention between monitoring the social situation and planning one’s own speech productions.” (Gould and Dixon, 1993). At the same time, turn length

Table 5: Correlation of age with dialogue act sequences

Label	$r(436)$	p	p_1	p_2
1. Statement-opinion (`sd`) + $\|B\|$ + Backchannel (`b`)	-0.184	0.00011	0.0081	0.0054
2. Statement-opinion (`sd`) + $\|E\|$ + Appreciation (`ba`)	0.182	0.00012	0.0075	0.0056
3. $\|B\|$ + Tag question (statement and YN question) (`qy^g`) …………… + $\|E\|$ + Negative non-no answer (`ng`)	0.181	0.00014	0.0108	0.0072
4. $\|B\|$ + Backchannel (`b`) + $\|E\|$ + YN question (`qy`)	-0.180	0.00015	0.0097	0.0152
5. $\|B\|$ + Action-directive (`ad`) + $\|E\|$	0.176	0.00022	0.0280	0.0162
6. Hedge (`h`) + Statement-opinion (`sd`)	-0.175	0.00023	0.0369	0.0301
7. $\|E\|$ + Backchannel (`b`) + $\|B\|$	0.175	0.00024	0.0328	0.0294

in number of words and dialogue acts produced by older speakers are also higher on average, as reported by e.g. (James et al., 1998) and confirmed with the Switchboard corpus data. It is then reasonable to expect that long turns elicit more backchannels by the dialogue partner in order to signal continued attention.

Correlation analysis has the advantage that the whole age range is modelled as opposed to binning interlocutors in an ‘old’ and ‘young’ group according to a more or less arbitrary ranges. Combined with randomly split data for identifying and testing hypotheses as well as combatting type I errors, prominent hypotheses were highlighted. However, it does not reveal whether interlocutors act differently when they are speaking to addressees of different ages, which is an interesting aspect to further investigate. To this purpose, speakers could be binned into ‘young’ and ‘old’ ($>$ 36 yrs) to allow for four groups: young-young, young-old and old-young, and old-old.

6 Conclusions & Future work

Dialogue act production of interlocutors across conversations showed that older interlocutors generally use fewer dialogue acts related to communication management (with more dialogue acts related to task and task management). Even though younger interlocutors produce more backchannels, older interlocutors use backchannels more at the start of a turn. They also tend to start their turn more often by repeating parts of what the speaker said before continuing with an opinion statement.

Current work, about to be completed, is addressing whether interlocutors act differently when they are speaking to addressees of different ages.

Future work in line of this study will investigate whether turn length affects general production rates. Furthermore, temporal aspects will be explored by linking the timing of the speech in the Switchboard I corpus to the dialogue act tags in the Switchboard Dialog Act Corpus. Recent research with the same dataset suggests that various social variables, including age, correlate significantly with turn-taking behavior (Grothendieck et al., 2009) and motivates further analysis.

Additionally, the use of language models as a way to capture relevant and possibly interesting subdialogues will be refined by using grammar induction (by unsupervised machine learning) to capture more complex re-occurring interaction patterns such as clarification or repairs subdialogues (Alexandersson and Reithinger, 1997; Geertzen, 2009).

References

Jan Alexandersson and Norbert Reithinger. 1997. Learning dialogue structures from a corpus. In *Proceedings of Eurospeech 1997*, page 22312234, Rhodes, Greece, September.

James Allen and Mark Core. 1997. Draft of DAMSL: dialog act markup in several layers. Unpublished manuscript.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26.

Anne Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34:351366.

Carlo E Bonferroni. 1935. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60, Rome, Italy. Tipografia del Senato.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

Deborah M. Burke and Meredith A. Shafto. 2004. Aging and language production. *Current directions in psychological science : a journal of the American Psychological Society*, 13(1):21–24. PMID: 18414600 PMCID: 2293308.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK, May.

Mark G. Core and James F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In David R. Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, page 2835, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Fredrik A Dahl, Margreth Grotle, Jūratė Šaltytė Benth, and Bård Natvig. 2008. Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain. *European journal of epidemiology*, 23(4):237–242.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Jeroen Geertzen. 2009. Dialogue act prediction using stochastic context-free grammar induction. In *Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference*, CLAGI '09, page 715, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kallirroi Georgila, Maria Klara Wolters, Vasilis Karaiskos, Melissa Kronenthal, Robert H. Logie, Neil Mayo, Johanna D. Moore, and Matthew Watson. 2008. A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

John Godfrey, Edward Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-92)*, page 517520, San Francisco, USA.

Odette N. Gould and Roger A. Dixon. 1993. How we spent our vacation: Collaborative storytelling by young and old adults. *Psychology and Aging*, 8(1):10–17, March.

J. Grothendieck, A. Gorin, and M. Borges. 2009. Social correlates of turn-taking behavior. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, pages 4745–4748.

Michel Hupet, Yves Chantraine, and Franois Nef. 1993. References in conversation between young and old normal adults. *Psychology and Aging*, 8(3):339–346, September.

L E James, D M Burke, A Austin, and E Hulme. 1998. Production and perception of "verbosity" in younger and older adults. *Psychology and Aging*, 13(3):355–367, September. PMID: 9793112.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, Institute of Cognitive Science, University of Colorado, USA.

Susan Kemper, Andrea FinterUrczyk, Patrice Ferrell, Tamara Harden, and Catherine Billington. 1998. Using elderspeak with older adults. *Discourse Processes*, 25(1):55–73.

Linda Mortensen, Antje S. Meyer, and Glyn W. Humphreys. 2006. Age-related effects on speech production: A review. *Language and Cognitive Processes*, 21(1-3):238–290, January.

Dana R Murphy, Meredyth Daneman, and Bruce A Schneider. 2006. Why do older adults have difficulty following conversations? *Psychology and aging*, 21(1):49.

Engagement driven Topic Selection for an Information-Giving Agent

Nadine Glas
Institut Mines-Télécom,
Télécom ParisTech, CNRS, LTCI
`glas`
`@telecom-paristech.fr`

Ken Prepin
Rakuten Institute
of Technology - Paris
`ken.prepin`
`@priceminister.com`

Catherine Pelachaud
CNRS, LTCI,
Télécom ParisTech
`pelachaud@`
`telecom-paristech.fr`

Abstract

We propose a model for conversational agents to select the topic of interaction in agent-initiated information-giving chat. By taking into account the agent's dynamically updated perception of the user's engagement, the agent's own preferences and its associations between topics, the agent tries to select the topic that maximises the agent and user's combined engagement. The model offers engagement driven dialogue management on the topic level.

1 Introduction

Conversational agents often employ a strict task-oriented dialogue structure in order to achieve the particular task for which they are built. Chat-based systems on the other hand, allow for less rigid interaction but the agent has less control of the topic of the interaction. Some applications however, ask for dialogue that falls in between these categories: Where there is not a clear task to achieve and where the interaction is not completely open either, but where there is freedom of topic choice within a certain domain. We are interested in the latter category, more specifically in interaction that is not task-driven but instead driven by social variables of the interaction. In this work, we present a topic selection model for a conversational agent, driven by the social variable *engagement*.

By taking into account the agent's perceived (detected) level of user engagement as well as the agent's own preferences and associations in the selection of a topic, we do not consider the dialogue merely from a user-oriented system point of view, but consider the agent as an interaction participant with human-like features that contributes to the interaction from its own point of view. In Section 4 we will detail the exact interpretation of these variables.

We consider engagement as *"the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction"* (Poggi, 2007). In order to favour the user's engagement level previous research manipulated the agent's non-verbal behaviour including gaze (Peters, 2005), gestures (Sidner and Lee, 2003), postures (Peters, 2005) and facial displays (Bohus and Horvitz, 2009), as well as the agent's verbal behaviour including the form (Glas and Pelachaud, 2014) and prosody (Foster, 2007) of its dialogue strategies. As mentioned above, certain interaction types however, also allow for an adaptation regarding the content of the agent's dialogue strategies. In this work we focus on the latter, by proposing a model where the agent initiates discussion topics that are adapted to the user.

In the following section we will first further specify the type of interaction and topic we are looking at. In Section 3 we present related work and in Section 4 we introduce the variables that will be taken into account in the topic selection model. In Section 5 we present the topic selection model itself. In Section 6 we discuss its configurations and in 7 its implementation. Section 8 concludes our findings.

2 Information-Giving Chat

The work we describe in this paper is conducted in the context of the French project 'Avatar 1:1' that aims at developing a human-sized virtual agent playing the role of a visitor in a museum. The agent's task is to engage human users in one-to-one face-to-face interaction about the museum and some of its art objects with the objective to give the visitors information about these subjects. The choice of the exact subject is secondary: what matters is that some amount of cultural information is transferred. We refer to this type of interaction as an *information-giving chat* (as opposed to information-seeking chat (Stede and Schlangen, 2004)). Just as information-seeking chat (Stede and Schlangen, 2004), information-giving chat is distinguished by its more exploratory and less task-oriented nature, while still being more structured than general free conversation.

The information-giving chat that is modelled in this particular project is agent-initiated in order to increase the likelihood of understanding the user's contributions. Due to the limitations of our natural language understanding module it is also the agent who introduces (initiates) the topics in the interaction.

The notion of *topic* in interactions can mean different things (Brown and Yule, 1983). We define topic from a discourse perspective as *what is being talked about in a conversation* (Brown and Yule, 1983). In the context of the information-giving chat defined above, each topic refers to the discussion phase of an artwork in the museum (O'Donnell et al., 2001). Each topic is thus associated to a fragment of conversation (similar to Macias-Galinde et al. (2012)) consisting of at least 1 pair of agent-user turns. Subtopics are subfragments of these larger conversation fragments and discuss a particular aspect of the artwork. For example, the artist of the artwork or the historical period during which it was created.

3 Related Work

Some previously built virtual agent systems give their users the opportunity to directly select or reject the topics of interaction (Bickmore et al., 2011; Kopp et al., 2005), thereby adapting the content of the interaction to the user. However, these systems only offer the user a choice for certain information. They do not present a conversational virtual agent than can select interaction topics itself based on dynamic social variables in the interaction.

In order for the agent to be able to select appropriate interaction topics itself, it needs to dispose of a domain knowledge representation mapping to the possible discussion topics. Several dialogue systems dispose of some kind of representation of domain knowledge, developed for various modules such as natural language understanding (Milward and Beveridge, 2003), topic tracking (Carlson and Hunnicutt, 1996; Jokinen and Wilcock, 2012), question-answering (Agostaro et al., 2005), response generation (Pilato, 2011), surface realization (Milward and Beveridge, 2003), and the selection or generation of dialogue topics (Chakrabory et al., 2007; Macias-Galindo et al., 2012; Stede and Schlangen, 2004). We are interested in the latter where domain knowledge is organised as in such a way that it represents (potential) interaction topics.

The topic representations can be divided in specific task-oriented models (Chakrabory et al., 2007) and non task-oriented models. As information-giving chat has a less task-oriented structure (Section 2) we focus on the latter category. In this category Macias-Galindo et al. (2012) use a semantic relatedness mechanism to transition between conversational snippets in an agent that engages in chatty dialogue, and Stede and Schlangen (2004) use an ontology-like topic structure that makes the agent produce coherent topic follow-ups in information-seeking chat. However, these systems do not take into account the user's engagement during the different discussion phases (topics). They are merely oriented towards dialogue coherence and are therefore not sufficient for an optimisation of engagement by topic selection.

Song et al. (2009) and Adam et al. (2010) do take into account the user's interests or engagement level in that they decide when the agent should switch topic. The systems are in charge of the timing of a topic change. The new topics are then respectively extracted from the web or a topic structure. For the selection of the topics themselves the user's engagement or preferences are not taken into account.

By using some concepts of the models described above, we aim at building a topic structure in the agent's mind to retrieve dynamically, during human-agent information-giving chat, engaging interaction topics. Opposite to existing topic selection systems that have focused exclusively on dialogue coherence, our topics will be generated from an agent perspective: The topic structure is representing a part of the agent's knowledge, which is located within the agent's mind, and the agent's objective is to constantly favour engagement. As such the topic selection will include human-like features by taking into account the agent's dynamically updated perception of the user's engagement, the agent's preferences and the agent's associations with respect to the current topic of conversation. In the section below we define these variables.

4 Variables for Topic Selection

In order to select engaging discussion topics, the agent needs to be able to predict the *user's engagement level* during the discussion of sofar unaddressed topics (objects). For this we need to know if there are any underlying observable *preferences* that can help the agent collect indications with regard to its prediction of the user's engagement. We interpret a preference as *"a relatively stable evaluative judgement in the sense of liking or disliking a stimulus"* (Scherer, 2005). Since a topic of conversation in our interaction setting corresponds to the discussion of a particular artwork, we verified by means of a perceptive study if there exists a relation between the user's engagement level during the discussion of an artwork with a virtual agent, and the user's preference for the physical artwork that is discussed. Below we shortly describe this study (for details see Glas and Pelachaud (2015)).

4.1 User Preferences and Engagement: Perceptive Study

We simulated a small museum in our laboratory by hanging photos of existing artworks on the walls. The artworks were chosen as to vary in style and type of affect they might evoke. When the participant finished observing the artworks in a first room, the visit continued in the next room where the participant talked with Leonard, introduced as a virtual character who also visits the museum. In the interaction Leonard discussed the different artworks from the museum in a random order.

After the interaction we presented the participants a questionnaire in which we asked indirectly for the user's engagement level during the different discussion phases, corresponding to each separate discussion around a museum object. We also asked for the user's preferences of the physical artworks.

Analyses of the data collected from 33 participants (13 female, aged 19-58) regarding the randomly discussed artworks have shown amongst others that the user's preference for a museum object is significantly, positively correlated with the user's engagement (*wanting to to be together with Leonard* $p < 0.001$, $\tau = 0.50$; *wanting to continue the interaction* $p < 0.001$, $\tau = 0.52$) during the discussion of this object with a virtual agent.

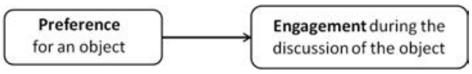

Preference for an object
Engagement during the discussion of the object

Figure 1: A human's preference and engagement.

From this finding we can derive that the user's preference for a physical object gives a direct indication of the user's engagement level during the discussion of this object (schematised in Figure 1). This makes that the characteristics (i.e. attributes) of a physical object can help the agent predict the user's future engagement level for the discussion of the object (further discussed in Section 4.3). The agent can then use its *predicted level of user engagement* for every object discussion to select an engaging topic of conversation.

4.2 Agent Preferences and Engagement

To represent human-like features in an agent that plays a museum visitor, the agent needs to have its own preferences for the artworks as well, as representing agent preferences is fundamental for any agent model (Casali et al., 2011). Besides, the preference representation of the agent can be used to express (consistent) agent appreciations, which has shown to significantly favour the user's engagement (Campano et al., 2015).

Following the correlation we found above (Section 4.1, Figure 1) an agent likes to talk most about its preferred topics as those maximise its own engagement. However, the agent we model also wants to engage the user. The agent thus tries to optimise the engagement level of both the user and the agent itself (from here onwards indicated as combined engagement). To achieve this, for each (sub)topic (object and characteristic) that can be addressed the agent calculates an *expected (predicted) level of combined engagement* and selects the one with the highest score as the next topic of discussion. In this way, the agent selects a new topic of conversation based on a combination of the agent's own preferences for the artworks and its prediction of the user's level of engagement during the discussion of the artworks. Figure 2 shows this relation.

4.3 Associations between Topics

A last human-like variable that needs to be represented when the agent selects a topic of conversation in information-giving chat are its own associations between topics. This is needed since events that share meaning or physical similarity become

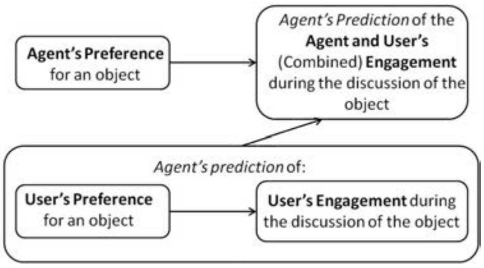

Agent's Preference for an object
Agent's Prediction of the Agent and User's (Combined) Engagement during the discussion of the object
Agent's prediction of:
User's Preference for an object
User's Engagement during the discussion of the object

Figure 2: The agent's prediction for the level of combined engagement during the discussion of an object depends on several variables.

associated in the mind (Dellarosa, 1988). Activation of one unit activates others to which it is linked, the degree of activation depending on the strength of association (Dellarosa, 1988). The discussion of one topic can thus be associated with other topics in the agent's mind by means of similarities or shared meanings between the topics. In the context of information-giving chat about museum objects each topic is revolving around an artwork. The topics can thus be associated in the agent's mind by similarities between the physical artworks. For example, an abstract painting by Piet Mondriaan may be associated with other abstract paintings, and/or with other works by Piet Mondriaan. In the topic selection model we will therefore represent the agent's associations by similarity scores between every pair of physical objects underlying two topics.

The associations (based on object similarities) allow the agent to make predictions about the user's engagement during sofar unaddressed topics: When the user has a certain engagement level during the discussion of the current topic (and thereby a related preference towards the current object under discussion (Section 4.1)), similar conversation topics are expected to have similar levels of user preference and are thus expected to lead to similar levels of user engagement (Figure 1). The topic selection model described in the following Section ensures that when the agent's predicted user engagement level for an associated topic is high enough (in combination with the agent's own preferences) it is a potential new topic of conversation, triggered by the agent's associations.

5 Topic Selection Model

In the spirit of (Stede and Schlangen, 2004) we define an ontology-like model of domain knowledge holding the conceptual knowledge and dialogue history. The model is part of the agent's knowledge and dynamically enriched with information representing the variables described above (Section 4).

The topics all consist of artwork discussions and are therefore not hierarchically ordered but represented in a non-directed graph $\{Obj, Sim\}$ (e.g. Figure 3) where each node represents an object among N objects: $\{Obj_i, i \in [1-N]\}$.

Each object node contains the object's name (corresponding to a topic) and its characteristics (attributes) that map to the topic's subtopics (see Section 2): $\{Char_n(Obj_i), n \in [1-C], i \in [1-N]\}$, where C is the number of characteristics for any object Obj_i.

All the topics are connected to each other by similarity scores: $\{Sim(Obj_i, Obj_j), i, j \in [1-N, i \neq j]\}$ (ranging from 0 to 1), which are responsible for the possible associations of the agent. Likewise, all the subtopics (characteristics of the objects) are connected to each other: $\{Sim(Char_n(Obj_i, Obj_j), i, j \in [1-N, i \neq j]\}$.

For every object and characteristic the agent has its own preferences: $\{Pref_a\ (Obj_i), Pref_a(Char_n(Obj_i)), a = agent\}$, where 0 corresponds to no liking and 1 to a maximum liking, following the definition in Section 4. The agent also has for every object and characteristic a continuously updated predicted level of the user's engagement during the discussion of these objects and characteristics at time $t+1$: $\{Eng^*_u(t+1, Obj_i), Eng^*_u(t+1, Char_n(Obj_i)), u = user\}$, where 0 refers to the minimum level of engagement to continue an interaction and 1 refers to the maximum level of engagement.

The latter two variables lead to a continuously updated predicted level of combined (user and agent) engagement by the agent for each object and characteristic for time $t+1$: $\{Eng^*_{u+a}(t+1, Obj_i), Eng^*_{u+a}(Char_n(t+1, Obj_i)\}$, ranging from 0 to 1. See Figure 3 for a graphical representation of the topic structure that incorporates all the variables of the (sub)topics.

For $\forall Obj_i, i \in [1-N]$ the agent's predicted level of combined engagement at time $t+1$ (described in Section 4.1) during the discussion of any Obj_i is:

$$\begin{aligned} &Eng^*_{a+u}(t+1, Obj_i) = \\ &w(t) \cdot Pref_a(Obj_i) + \\ &(1-w(t)) \cdot Eng^*_u(t+1, Obj_i) \end{aligned} \tag{1}$$

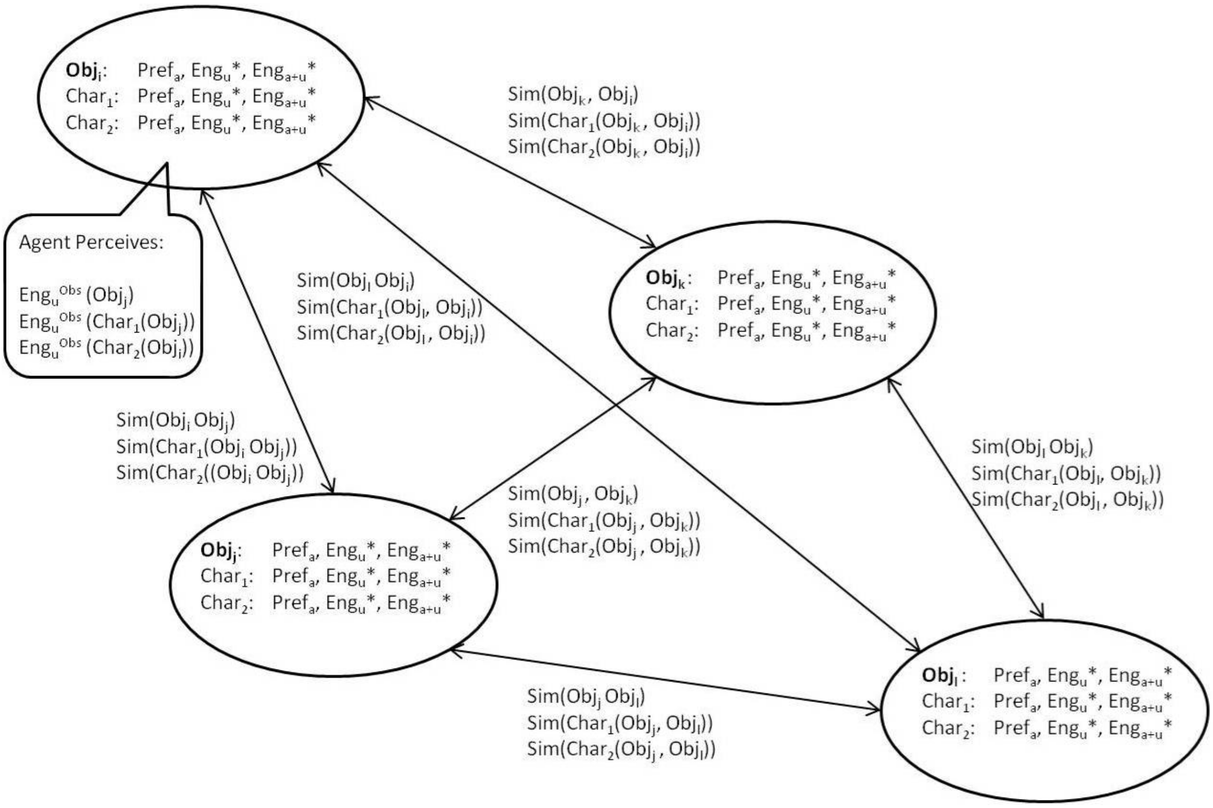

Obj_i: Pref_a, Eng_u*, Eng_a+u*
Char_1: Pref_a, Eng_u*, Eng_a+u*
Char_2: Pref_a, Eng_u*, Eng_a+u*
Agent Perceives:
Eng_u^Obs (Obj_j)
Eng_u^Obs (Char_1(Obj_j))
Eng_u^Obs (Char_2(Obj_j))
Sim(Obj_k, Obj_i)
Sim(Char_1(Obj_k, Obj_i))
Sim(Char_2(Obj_k, Obj_i))
Obj_k: Pref_a, Eng_u*, Eng_a+u*
Char_1: Pref_a, Eng_u*, Eng_a+u*
Char_2: Pref_a, Eng_u*, Eng_a+u*
Sim(Obj_l Obj_i)
Sim(Char_1(Obj_l, Obj_i))
Sim(Char_2(Obj_l, Obj_i))
Sim(Obj_i Obj_j)
Sim(Char_1(Obj_i Obj_j))
Sim(Char_2((Obj_i Obj_j))
Sim(Obj_l Obj_k)
Sim(Char_1(Obj_l, Obj_k))
Sim(Char_2(Obj_l, Obj_k))
Sim(Obj_j, Obj_k)
Sim(Char_1(Obj_j, Obj_k))
Sim(Char_2(Obj_j, Obj_k))
Obj_j: Pref_a, Eng_u*, Eng_a+u*
Char_1: Pref_a, Eng_u*, Eng_a+u*
Char_2: Pref_a, Eng_u*, Eng_a+u*
Obj_l: Pref_a, Eng_u*, Eng_a+u*
Char_1: Pref_a, Eng_u*, Eng_a+u*
Char_2: Pref_a, Eng_u*, Eng_a+u*
Sim(Obj_j Obj_l)
Sim(Char_1(Obj_j, Obj_l))
Sim(Char_2(Obj_j, Obj_l))

Figure 3: The graph representing an example of a topic structure in the agent's mind at a time t. Each circle represents a topic (object) where Obj_j is the current object under discussion.

Where $w(t)$ is the ratio indicating to what extent the agent values its own preferences in comparison to the user's engagement at moment t. The same equation holds for $\forall Char_n(Obj_i), n \in [1-C], i \in [1-N]$ by replacing Obj_i by $Char_n(Obj_i)$).

The agent's prediction of the user's future ($t+1$) engagement during the discussion of Obj_i (and its characteristics by replacing Obj_i by $Char_n(Obj_i)$)) is:

$$\begin{aligned} &Eng_u^*(t+1, Obj_i) = \\ &Eng_u^{obs}(t, Obj_j) \cdot Sim(Obj_j, Obj_i) + \\ &Eng_u^*(t, Obj_i) \cdot (1 - Sim(Obj_j, Obj_i)) \end{aligned} \tag{2}$$

Where $Obj_i \neq Obj_j$ and $Eng_u^{obs}(t, Obj_j)$ is the agent's observed level of user engagement during the discussion of Obj_j at time t.

5.1 Initial State

The initial state of the topic structure used for the agent's topic selection contains all the objects and characteristics that are known to the agent. For these entities dialogue fragments have been created. For $\forall Obj_i, i \in [1-N]$ and $\forall Char_n(Obj_i)$, $n \in [1-C], i \in [1-N]$ the agent's preferences and similarity scores can be initialised at any value between 0 and 1. In case we want agent associations that correspond to observable objective similarities between objects, theoretically defined similarity measures (e.g. Mazuel and Saboutret, 2008) can be used. In the latter case the similarity score between objects can be derived directly from the similarity scores of their characteristics. We further initialise for $\forall Obj_i, i \in [1-N]$:

$$\begin{aligned} &Eng_u^*(t_0+1, Obj_i) = Pref_a(Obj_i) \\ &\text{and } w(t_0) = 1 \end{aligned} \tag{3}$$

The same holds $\forall Char_n(Obj_i), n \in [1-C], i \in [1-N]$ (replacing Obj_i by $Char_n(Obj_i)$)). This makes that for time t_0+1 the predicted user engagement of every object and characteristic equal the agent's preferences. However, this assumption is only used as a starting point for future predictions of the user's engagement that will be based on observed user behaviour (Equation 2). At the start of the interaction (t_0), the agent only takes into account its own preferences, indicated by $w(t_0) = 1$. The first topic the agent introduces in the interaction is the one for which it predicts the highest level of combined engagement at time $t_0 + 1$:

$$max\{Eng^*_{a+u}(t_0 + 1, Obj_{1-n})\} \quad (4)$$

The agent introduces one by one the subtopics of this first topic for which:

$$\{Eng^*_{a+u}(t_0 + 1, Char_n(Obj_i))\} > e \quad (5)$$

Where e is a threshold for the minimum level of predicted mutual engagement level that the agent finds acceptable for the interaction. For example the agent can decide to only talk about the subtopics that are predicted to lead to half the maximum level of engagement, setting e to 0.5.

5.2 Updating

A new topic is selected when either: 1) The current topic is finished, meaning that the conversational fragment has been uttered completely. Or 2) the detected level of user engagement during an interval I within the discussion phase of an object is below a threshold z. The description of the user engagement detection method itself lies outside the scope of this paper. The required length of I and level of z, which determine when the user's engagement level (detected by the agent) should lead to a topic switch, will be studied in future work.

At any time t, just before selecting a new topic of interaction the agent first updates the weights in the topic structure with information that is gathered during the previous discussion phases. In the rest of this section we describe how.

For $\forall Obj_i, i \in [1 - N]$ that are part of the dialogue history (already discussed topics) we set:

$$Eng^*_u(t + 1, (Obj_i)) = 0 \text{ and } w(t) = 0 \quad (6)$$

Similarly for $\forall Char_n(Obj_i), n \in [1 - C], i \in [1 - N]$ that are part of the dialogue history. This implies that the agent makes the assumption that once a topic has been addressed the user does not want to address it again. The agent values this over its own preferences ($w(t) = 0$). This simplification makes that the system shall not discuss a topic twice.

$\forall Obj_i, i \in [1 - N]$ and $\forall Char_n(Obj_i), n \in [1 - C], i \in [1 - N]$ that are not in the dialogue history the agent's prediction for the user's engagement at time $t + 1$: $Eng^*_u(t + 1, Obj_i)$, as well as the agent's prediction for the agent and user's combined engagement level at time $t+1$: $Eng^*_{a+u}(t + 1, Obj_i)$ are updated by Equation 1 and Equation 2. This is done by entering the agent's detected (observed) overall level of user engagement during the discussion phases of the lastly discussed object and each of its characteristics (at time t): $Eng^{obs}_u(t, Obj_j)$ and $Eng^{obs}_u(Char_k(t, Obj_j))$. This update makes sure that the agent's observed user engagement level influences the predicted (user and combined) engagement levels for the objects and characteristics that the agent associates with the previously discussed (sub)topics. As mentioned before, the detection method of the user's engagement level lies outside the scope of this paper.

In circumstances where a detection of the user's engagement is not possible Equation 1 and Equation 2 can be updated by entering the user's explicitly uttered preferences for the lastly discussed object and characteristics at the place of respectively $Eng^{obs}_u(t, Obj_j)$ and $Eng^{obs}_u(Char_k(t, Obj_j))$. This follows from the finding that a user's preference is directly related to the user's engagement (Section 4.1).

5.3 Topic Selection

Whenever a new topic needs to be introduced (see previous section) it is selected in the same way as the first topic of the interaction:

$$max\{Eng^*_{a+u}(t + 1, Obj_{1-n})\} \quad (7)$$

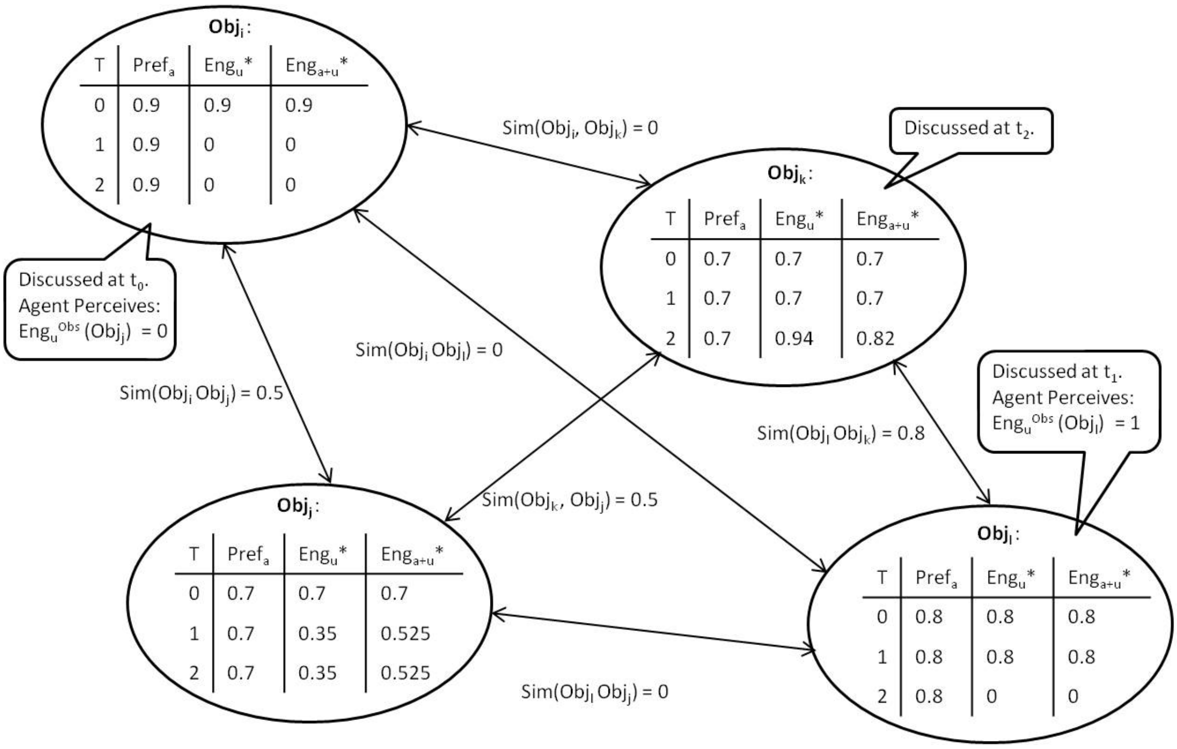
In this way, the agent tries to optimise the combined engagement. The selected subtopics of this topic are, like the first subtopics of the interaction, those where:

$$\{Eng^*_{a+u}(t + 1, Char_n(Obj_o))\} > e \quad (8)$$

5.4 Example

For the sake of clarity, in this Section we demonstrate the working of the topic selection model with a small example topic structure shown in Figure 4. In this example the agent knows about the 4 topics (objects) that are listed in Table 1.

Figure 4 shows how the variables for each topic can evolve over time t_{0-2} during an interaction. Due to space limitations we limit this example to the calculation and selection of topics. The calculation of the weights of the subtopics occurs in exactly the same manner. Only the ultimate selection of subtopics differs slightly as described in Section 5.3.



Obj_i:

T	$Pref_a$	Eng_u*	Eng_{a+u}*
0	0.9	0.9	0.9
1	0.9	0	0
2	0.9	0	0

Discussed at t_0.
Agent Perceives:
$Eng_u^{Obs}(Obj_j) = 0$

Obj_k:

T	$Pref_a$	Eng_u*	Eng_{a+u}*
0	0.7	0.7	0.7
1	0.7	0.7	0.7
2	0.7	0.94	0.82

Discussed at t_2.

Obj_j:

T	$Pref_a$	Eng_u*	Eng_{a+u}*
0	0.7	0.7	0.7
1	0.7	0.35	0.525
2	0.7	0.35	0.525

Obj_l:

T	$Pref_a$	Eng_u*	Eng_{a+u}*
0	0.8	0.8	0.8
1	0.8	0.8	0.8
2	0.8	0	0

Discussed at t_1.
Agent Perceives:
$Eng_u^{Obs}(Obj_l) = 1$

Figure 4: Example of the evolution of the weights in the topic structure over time t_{0-2}. In this example, at each t, $w = 0.5$.

Object	**Type**	**Artist**
Obj_i	Statue	Antiquity
Obj_j	Statue	17th century
Obj_k	Painting	17th century
Obj_l	Painting	18th century

Table 1: The objects of the example topic structure of Figure 4.

The values of the variables for each topic at time t_0 represent the initial state. Given that at this moment $Eng{*}_{a+u}$ is the highest for Obj_j, this topic is the first to be selected for discussion. When the agent then perceives a minimum level of user engagement during the discussion of this topic, the updated variables (t_1) lead to the selection of object Obj_l as next object, which has nothing in common with the former object. To show the opposite extreme situation, during the discussion of object Obj_l the agent perceives a maximum level of user engagement, leading to the selection of Obj_k as next topic, staying close to the characteristics of the former object. Of course Figure 4 is only a limited example and not sufficient to illustrate the full potential of the tradeoff between agent and user oriented variables in the selection of a topic.

6 Topic Selection Configurations

As described in Section 5.2 the preference, engagement and similarity weights in the topic structure can be initialized at any value ranging from 0 to 1. The freedom of initialising these variables as desired allows for different configurations. The initialisation of the agent's preferences, for instance, can reflect different types of agents (as recommended by Amgoud and Parsons (2002)) but can also be initialised, for example, at values that are close to the users' preferences in previous interactions. The agent's preferences for the objects can be directly related to the sum of its preferences for the characteristics of the object or not. It is also possible to attribute more importance to the preference for one characteristic than to another. The same holds for the similarity values. For example, to model an agent that is particularly focused on history, the similarity and preference weights of the characteristic "period" may have a larger impact on the similarities and preference of the entire object than the other characteristics of the object. The initialisation of the graph can be simplified with the help of a museum catalogue that already lists the objects and their characteristics.

The topic selection model can be easily extended to other domains that can be structured similarly as museum objects. This means that the agent needs to have its preferences for the different topics and can associate the topics to each other by means of similarity scores. Selecting subtopics as described in Section 5 is only possible if the topics' characteristics (attributes) can be defined.

7 Implementation and Dialogue Management

For the management of the multimodal behaviour of the agent we use the hierarchical task network Disco for Games (Rich, 2012) that calls prescripted FML files, which are files that specify the communicative intent of an agent's behaviour and include the agent's speech (Heylen et al., 2008). As Disco is developed for task-oriented interactions it offers a fixed, scripted order of task execution where agent contributions and user responses alternate. As mentioned in Section 2, in our project each (sub)topic is associated to a scripted fragment of conversation, consisting of 1 or multiple pairs of agent-user turns. Within such a conversation fragment we can thus directly use the Disco structure. The agent executes the tasks that consist of talking about the object while the local responses of the user drive the dialogue further in the network.

In between (sub)topics it is different. In information-giving-chat we cannot foresee, and thus predefine, the topics and their order of discussion as they are selected by the agent during the interaction. Therefore, at any time a topic switch is required, we overwrite the fixed task structure proposed by Disco by calling from an external module the appropriate tasks that map to the selected (sub)topics. The external module is the topic selection module of the agent. In this way we continuously paste in real time dialogue parts to the ongoing conversation.

This procedure makes that local dialogue management is controlled by Disco and topic management is controlled by the external topic selection module, thereby adding flexibility to the existing task-oriented system, resulting in a more adaptive and dynamic dialogue. The agent's topic selection module could be connected to any other task and/or dialogue system as well.

8 Conclusion and Future Work

In this work we have proposed an engagement driven topic selection model for an information-giving agent. The model avoids the need for any pre-entered information of the user. Instead, it dynamically adapts the interaction by taking into account the agent's dynamically updated perception of the user's level of engagement, the agent's own preferences and its associations. In this way, the interaction can be adapted to any user. The model's configurations also allow for different types of agents. By connecting the topic selection model to existing task-oriented systems we proposed a way to construct a dynamic interaction where the agent continuously pastes in real time dialogue parts ((sub)topics) to the ongoing conversation.

In the future we would like to perform a perceptive study to evaluate the topic selection model in human-agent interaction. However, for this we first need to plan some additional research.

First, we will study the different ways of switching topic on a dialogue generation level. Even when two consecutive topics have no characteristics in common, the agent needs to present the new topic in a natural way in the conversation without loosing the dialogue coherence (Levinson, 1983). Strategies that we will consider to achieve this include transition utterances by the agent that make the agent's associations explicit, transition utterances that recommend interesting (e.g. similar/opposite) artworks to the user, and transition utterances that refer to the artworks' locations within the museum.

Further, before evaluating the topic selection module in interaction with users, we will also have a closer look at the timing of topic transitions. We noted that a topic switch is needed, amongst others, when the current topic leads to a very low user engagement during a certain time interval. We will need to determine the exact interval of the engagement detection that is required and thereby determine the timing of a possible topic transition.

Possible extensions to the topic selection model we would like to explore in the future include the option of coming back to previously addressed topics, considering agent preferences that may change during the interaction and dealing with similar information-giving conversational fragments for different topics.

Acknowledgments

We would like to thank Charles Rich for his help with Disco. This research is partially funded by the French national project Avatar 1:1, ANR MOCA, and the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR, within the Investissements d'Avenir program under reference ANR-11-IDEX-0004-02.

References

Carole Adam, Lawrence Cavedon, and Lin Padgham. 2010. Flexible conversation management in an engaging virtual character. *International Workshop on Interacting with ECAs as Virtual Characters.*

Francesco Agostaro, Agnese Augello, Giovanni Pilato, Giorgio Vassallo, and Salvatore Gaglio. 2005. A conversational agent based on a conceptual interpretation of a data driven semantic space. *Advances in Artificial Intelligence*, 381–392.

Leila Amgoud and Simon Parsons 2002. Agent dialogues with conflicting preferences. *Intelligent Agents VIII*, 190–205.

Timothy Bickmore and Julie Cassell. 1999. Small talk and conversational storytelling in embodied conversational interface agents. *AAAI fall symposium on narrative intelligence*, 87-92.

Timothy Bickmore and Julie Cassell. 1999. Small talk and conversational storytelling in embodied conversational interface agents. *AAAI fall symposium on narrative intelligence*, 87–92.

Timothy Bickmore, Daniel Schulman and Langxuan Yin. 2010. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, 24(6), 648–666.

Timothy Bickmore, Laura Pfeifer and Daniel Schulman. 2011. Relational agents improve engagement and learning in science museum visitors. *Intelligent Virtual Agents*, 55–67.

Cynthia Breazeal and Brian Scassellati. 1999. How to build robots that make friends and influence people. *Proceedings of the International Converence on Intelligent Robots and Systems 2*, 858–863.

Gillian Brown and George Yule. 1983. Discourse Analysis. *Cambridge University Press.*

Dan Bohus and Eric Horvitz. 2009. Models for multiparty engagement in open-world dialog. *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 225–234.

Sabrina Campano, Chlo Clavel, and Catherine Pelachaud. 2015. I like this painting too: When an ECA Shares Appreciations to Engage Users. *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1649–1650.

Ana Casali, Llus Godo and Carles Sierra. 2011. A graded BDI agent model to represent and reason about preferences. *Artificial Intelligence*, 175(7), 1468-1478.

Rolf Carlson and Sheri Hunnicutt. 1996. Generic and domain-specific aspects of the Waxholm NLP and dialog modules. *Proceedings of the 4th International Conference on spoken Language Processing*, 677–680.

Sunandan Chakraborty, Tamali Bhattacharya, Plaban K. Bhowmick, Anupam Basu, and Sudeshna Sarkar. 2007. Shikshak: An intelligent tutoring system authoring tool for rural education. *International Conference on Information and Communication Technologies and Development*, 1–10.

Denise Dellarosa. 1988. The psychological appeal of connectionism. *Behavioral and Brain Sciences*, 11(01), 28–29.

Mary Ellen Foster. 2007. Enhancing human-computer interaction with embodied conversational agents. *In Universal Access in Human-Computer Interaction. Ambient Interaction*, 828–837.

Nadine Glas and Catherine Pelachaud. 2014. Politeness versus Perceived Engagement: an Experimental Study. *Proceedings of the 11th International Workshop on Natural Language Processing and Cognitive Science*, 135–147.

Nadine Glas and Catherine Pelachaud. 2015. User Engagement and Preferences in Information-Given Chat with Virtual Agents. *Workshop on Engagement in Social Intelligent Virtual Agents*, Forthcoming.

Fritz Heider. 1958. The Psychology of Interpersonal Relations. *Psychology Press.*

Dirk Heylen, Stefan Kopp, Stacy C. Marsella, Catherine Pelachaud, and Hannes Vilhjlmsson. 2008. The next step towards a function markup language. *Intelligent Virtual Agents*, 270–280.

Kristiina Jokinen and Graham Wilcock. 2012. Constructive Interaction for Talking about Interesting Topics. *Proceedings of Eighth International Conference on Language Resources and Evaluation*, 404-410.

Souhila Kaci. 2011. Working with Preferences. Less Is More. *Springer.*

Stefan Kopp, Lars Gesellensetter, Nicole C. Krmer and Ipke Wachsmuth. 2005. A conversational agent as museum guide-design and evaluation of a real-world application. *Intelligent Virtual Agents*, 329–343.

Stephen C. Levinson. 1983. Pragmatics. *Cambridge textbooks in linguistics.*

Daniel Macias-Galindo, Wilson Wong, John Thangarajah and Lawrence Cavedon. 2012. Coherent topic transition in a conversational agent. *Proceedings of the 13th Annual Conference of the International Speech Communication Association*, 1–4.

Laurent Mazuel and Nicolas Sabouret. 2008. Semantic relatedness measure using object properties in an ontology. *Springer Berlin Heidelberg*, 681–694.

David Milward and Martin Beveridge. 2003. Ontology-based dialogue systems. *Proceedings of the 3rd Workshop on Knowledge and reasoning in practical dialogue systems*, 9–18.

Mick O'Donnel, Chris Mellish, Jon Oberlander and Alistair Knott. 2001. ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3) : 225–250.

Christopher Peters. 2005. Direction of attention perception for conversation initiation in virtual environments. *Proceedings of the 5th International Conference on Intelligent Virtual Agents*, 215–228.

Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. Engagement capabilities for ECAs. *AAMAS'05 workshop Creating Bonds with ECAs*.

Giovanni Pilato, Agnese Augello, and Salvatore Gaglio. 2011. A modular architecture for adaptive ChatBots. *Fifth International Conference on Semantic Computing*, 177–180.

Isabella Poggi. 2007. Mind, hands, face and body: a goal and belief view of multimodal communication. *Weidler*.

Charles Rich and Candace L. Sidner. 2012. Using collaborative discourse theory to partially automate dialogue tree authoring. *Intelligent Virtual Agents*, 327-340.

Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social science information*, 44(4), 695–729.

Candace L. Sidner and Christopher Lee. 2003. An architecture for engagement in collaborative conversations between a robot and humans. *Mitsubishi Electric Research Labs TR2003-13*.

Xin Song, Kazuki Maeda, Hiroyuki Kunimasa, Hiroyuki Toyota, and Dongli Han. 2009. Topic control in a free conversation system. *International Conference on Natural Language Processing and Knowledge Engineering*, 1–6.

Manfred Stede and David Schlangen. 2004. Information-seeking chat: Dialogues driven by topic-structure. *Proceedings of the 8th Workshop on Semantics and Pragmatics of Dialogue*, 117–124.

Building and Applying Perceptually-Grounded Representations of Multimodal Scene Descriptions

Ting Han **Casey Kennington** **David Schlangen**
CITEC / Dialogue Systems Group / Bielefeld University
`first.last@uni-bielefeld.de`

Abstract

"You see a red building, and then behind that [gesture] you turn left". Hearing this kind of route description, only to apply its instructions at a later time, is a difficult task. The content of the description has to be memorised, and then, when the time comes to make use of it, be applied to the present situation. This makes for a good test case for a model of situated dialogue understanding, as the accuracy and applicability of a constructed multimodal content representation can be directly tested. In this paper, we present a model of a simplified version of this general task (namely, describing spatial scenes) and discuss three variants for realising the 'extraction', memorisation, and application of the content of route descriptions. We evaluate the approach and the variants with an implementation of the model, using a corpus of descriptions and application situations.

1 Introduction

Describing routes to destinations not currently in view, then understanding and later following such descriptions, is among the hardest language-related tasks (Schneider and Taylor, 1999). The description giver must imagine to herself the spatial layout of the scene through which the route leads, must imagine movement through that scene, and then encode all this in speech and gestures. The recipient in turn must represent to himself the content of the description, in such a way that it can later indeed form the basis for navigation.

In this paper, we model the task of the *recipient* of such a description, in a somewhat simplified version. Being presented with a multimodal description of a spatial configuration of landmarks, such as illustrated by (1) and gestures as shown in Figure 1 below, we assume that the recipient builds a representation of the content of this description and is then later able to use this representation to recognise the described scene in a set of candidate scenes.

(1) There is a red circle here [gesture] and slightly above it [gesture] a blue L.

We propose, compare and explore a range of different models for building and applying such representations, which we implement in a dialogue processing system and evaluate on a corpus of scene descriptions.

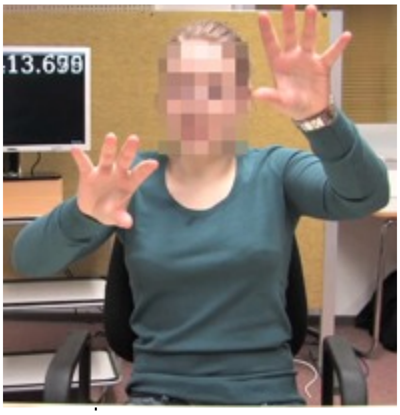


Figure 1: Providing a multimodal scene description

Our desiderata for these representations are as follows: That they represent equally the contribution by language and by gesture (which goes beyond what formal semantics-based approaches typically do); that the mapping from lexical entries to non-logical constants is well-motivated (where in contrast in formal semantics what should be arbitrary symbols is often suggestively named, e.g. `red` as translation of the word *red*); and that these non-logical constants are perceptually grounded (and not just equiped with a model-theoretic interpretation that simply states the extension).

We describe in general terms the structure of the representations in the following section, along with the variants within that structure that we explore. In Section 3 we further specify the modelling task and describe how we implemented it. The implementation is then evaluated in Section 4. We then discuss related literature and conclude.

2 Representing Scene Descriptions

Descriptions such as illustrated by example (1) form a type of mini-discourse, where referents for objects are introduced into the discourse and constraints are added. The basic structure of the representation format we use is consequently inspired by Discourse Representation Theory (Kamp and Reyle, 1993), as can be seen by the representation (schema) for the example given in (2).

(2)

o_1, g_1, o_2, g_2
o_1: *transl(*red circle*)* $g_1 : (x_1, y_1)$ $pos(o_1, \phi(g_1))$ $slightly\_above(o_1, o_2)$ o_2: *transl(*blue L*)* $g_2 : (x_2, y_2)$ $pos(o_2, \phi(g_2))$

The gestural component is represented by the "gesture referents" g_1 and g_2, and we simplify in assuming that they only contribute a single point in space (the position of the stroke of the deictic gesture). The connection between the verbal content and the gestural content is indicated by predicates stating that the gestures, respectively, specifiy the positions of the objects.[1]

We explore three different ways of filling in the details that are glossed over in (2) with the function $transl()$, which is supposed to translate from the utterances to its logical form.

- In **Variant A**, we translate the referring expressions simply into a sequence of lemmata. This would lead to a representation of "red circle" as $red, circle$.

- In **Variant B**, the translation proceeds by specifying a semantic frame (Fillmore, 1982), but here by way of more practically-oriented approaches to spoken language understanding (Tur and De Mori, 2011)) for object descriptions, leading to, for example $\begin{bmatrix} shape : circle' \\ colour : red' \end{bmatrix}$ for "red circle". This presupposes availability of a process that can do such a mapping; e.g., a lexicon that links lexical items and such frame elements, and a prespecified repertoire of attributes and values for them.

- In **Variant C** finally, we map the referring expression into a sequence of symbols (similar to Variant B) where however the repertoire of these symbols comes from an automatic learning process, and thus does not necessarily correspond to pre-theoretic notions of the meaning of such attributes.

All variants have in common that the symbols used in the representation are perceptually grounded, that is, their applicability in a given context can be determined by representing that context through perceptual (here, visual) features.

To make these proposals more concrete, we put them to use in a specific application, which will be described in the next section.

3 Processing Scene Descriptions

3.1 The Scene Retrieval Task

The specific task that we are modelling is the following: Given – in real time, word by word – a verbal/gestural description as in (1), construct a representation of the relevant content. Then, when the representation is built, use it to identify in a set of visually presented scenes the one that best conforms to the description. The task hence requires a) *constructing* the representation, based on perceived speech and gestures, and b) *applying* it in a (later) visually perceived context. Performance on the retrieval task gives a practical measure for the quality of the representation; if the representation does indeed capture the relevant content, it should form the basis for identifying that what was described.

Figure 2 shows an example scene which we used in our evaluation experiments. In all of the scenes, there are three puzzle pieces (more precisely, pentomino pieces constructed out of 5 squares in different configurations, which leads to 12 possible shapes, from which three are randomly

---

[1] Following Lascarides and Stone (2009), we assume that they do this via a context-specific function that maps the positions in gesture space to the intended real-world positions; but we do not further develop this part here.

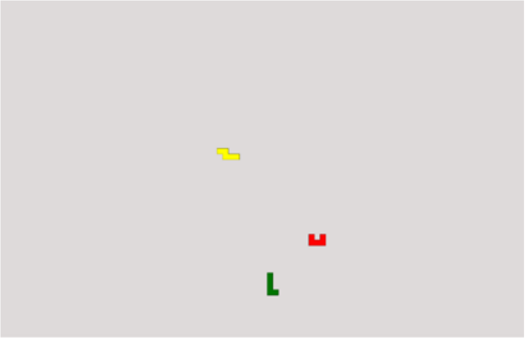


Figure 2: Scene example

selected), with a randomly determined color and position.

3.2 The Processing Pipeline

The verbal/gestural description of a scene is processed by a processing pipeline as illustrated in Figure 4. As shown in the figure, the system take speech and gesture as input. Speech is processe by an ASR which produces output word-by-wor The output then is fed into a segmentation mo ule that decides when a new object is introduce in the discourse. In parallel, a motion capture se sor records hand motion data and sends the dat to a deictic gesture detector. This detector sends signal to the segmentation module when a deicti gesture is detected. The segmentation module i tegrates the deictic gesture with the correspondin object information. The integrated information then sent to a representation module which buil the representation for the incoming description.

At a later time, and after the full description h been perceived, it is used to make a decision in th retrieval task. The scenes among which the d scribed one is to be found are given (as image to a computer vision module, which recognis the objects in the scenes and computes a featu vector for each, containing information about the colour of the object, the number of edges, its skewness, position, etc.; i.e., crucially, the object is not represented by a collection of symbolic property labels, but by real-valued features. The application module takes this representation for each candidate scene as input, and computes a score for how well the stored representation of the description content matches the candidate scene. For this, it makes use of the perceptually-grounded nature of the symbols used in the content representation, which connect these with the object feature vectors. The scene with the highest score finally is chosen as the one that is retrieved.

In the following, we describe some of these processing steps in more detail.

3.3 Applying Gestural Information

As described above, the discourse representation includes information about positions indicated by the gestures. To make use of this information in distinguishing between scenes, the first step is to compute for each scene the likelihood that it, with the position that objects are in, gave rise to the observed (and represented) gesture positions. This is not as trivial as it may sound, as the gesture positions are represented in a coordinate system given by the motion capture system, whereas the object positions are relative to the image coordinate system. Moreover, the gestures may have been performed sloppily. Finally, on a more technical level, the labels that the segmentation module as-

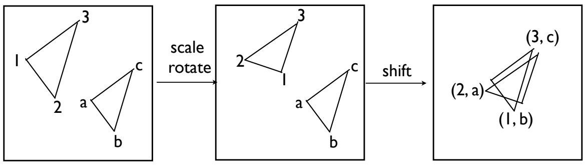


Figure 3: Example of a good mapping (top) and bad mapping (bottom), numbered IDs represent the perceived objects, the letter IDs represent the described objects.

To address the latter question (which description objects to compare with which computer vision object), we simply try all permutations of mappings. For each mapping a score is then computed for how well the gestured configuration under a given mapping can be transformed into the scene configuration. This is illustrated in Figure 3. First, the gestured configuration is projected into the same coordinate system as the scene configuration, and then it is scaled, rotated and shifted to be as congruent with the scene configuration as possible. In Figure 3, where the top target mapping between description object IDs and scene ob-

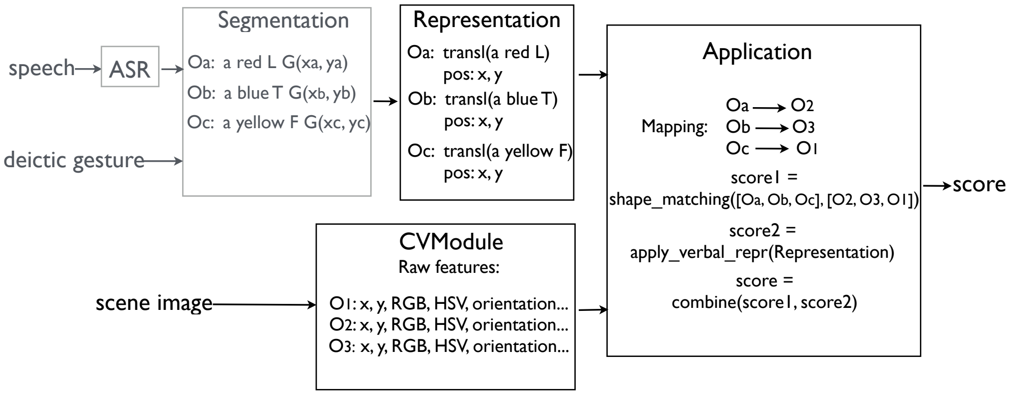


Figure 4: Processing pipeline

ject IDs is sensible, this operation leads to a good fit, the bottom mapping is not as good. (It will be even worse when attempting to map a gesture configuration into a scene configuration that is wildly different; e.g. a triangle into a sequence of objects placed in one line.) We assume that we have available a model trained on observed gestures for known positions, which can turn this distance score into a probability (i.e., the likelihood of this gesture configuration being observed when the scene configuration is the intended one).

On a technical level, this works as follows. The positions of the three objects in the description and in the visual scene can be represented as matrices S_d and S_v of the form:

$$S = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \end{pmatrix} \tag{1}$$

With a set of parameters p

$$p = [\theta,\, t_x,\, t_y,\, s] \tag{2}$$

where θ is the rotating angle; t_x (t_y) stands for the shift value on the x (y) axis; s is the scaling parameter. We scale, rotate and shift matrix S_v to get a transformed matrix S_t:

$$S_t(x, y) = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + s \begin{pmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{3}$$

By minimizing the cost function:

$$E = min \parallel S_t - S_d \parallel \tag{4}$$

we compute the optimal p. The distance between the resulting optimal S_t and S_d gives a metric for the goodness of the result, which is the input for a likelihood model that turns this into a probability.

3.4 Knowledge from Prior Experience

We assume that our system brings with it knowledge from previous experience with object descriptions. This knowledge is used (at least in some variants) for the task of mapping to logical form, and in all variants for the perceptual grounding of the symbols in the logical form. In what follows, we first briefly describe the corpus of interactions from which this prior knowledge is distilled.

3.4.1 The corpus: TAKE

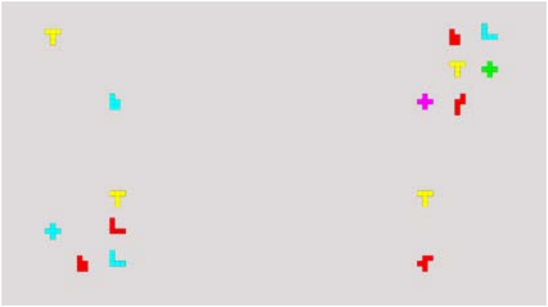


Figure 5: Example TAKE scene used for training.

As source for this knowledge, we use the TAKE corpus (Kousidis et al., 2013). In a Wizard-of-Oz study, participants were presented on a computer screen with a scene of pentomino pieces (as in Figure 5) and asked to identify one piece to a "system" by describing and pointing to it. The utterances, arm movements, scene states and gaze information were recorded as descibed in Kousidis et al. (2012). In total, 1214 episodes were recorded from 8 participants (all university students). The corpus was further processed to include raw visual features (such as color, shape HSV, RGB values etc.) of each pento tile for each scene (in

	DESCRIPTION	REPRESENTATION	APPLICATION	VISUAL INPUT
A	speech + gesture	word stems	word classifiers	raw visual object and scene features
B		property labels	property classifiers	
C		cluster labels	cluster classifiers	

Table 1: Overview of variants.

the same way as the computer vision module described above does); it also includes for each object symbolic properties (e.g., *green*, *X* (a shape)), and for the intended referent the utterance that the participant used to refer to it.

3.4.2 (Learning) Mappings to Logical Form

As described above, one difference between the variants of our model lies in how they realise the $transl()$ function from representation (2). Only variant B actually uses the data to learn this mapping, but we describe all variants here. In all variants, there is a preprocessing step that normalises word forms (by *stemming* them using NLTK (Loper and Bird, 2002)). This will map for example all of *grün, grünes, grüne* into *grun*. Thus, this step reduces the size of the vocabulary that needs to be mapped.

Variant A For variant A, stemming is all that is done in terms of mapping into logical form, and an object description is translated into the sequence of its stemmed words.

Variant B For variant B, similar to the model presented in (Kennington et al., 2013), we learn a simple mapping from words to symbolic property labels, based on co-occurrence in the training data. (E.g., we will have observed that the word *green* occured when the referent had the property `green`, strengthening the link between that word and that property.) The model gives us for each word a probability distribution over properties; we chose the most likely property (averaging over the contribution of all words) as the representation for the description. Note that this variant does not require a pre-specified lexicon linking words to properties, but it does require a pre-specified set of properties (e.g., `green`, `red`, etc., totaling 7 colour and 12 shape properties).

Variant C We overcome the latter limitations (pre-specifying properties) in this variant. As will be described below, for variant A we learn for each word (stem) a classifier that links it to perceptual input. These classifiers themselves can be represented as vectors (the regression weights of the logistic regression, see below). Using the intuition that words with similar meaning should give rise to similarly behaving classifiers (e.g., the classifier to “light green” should respond similarly–not identically–to that for “green”), we ran a clustering algorithm (k-means clustering) on the set of classifier vectors. The resulting clusters, through their centroids, can then themselves be turned again into classifiers. This effectively reduces the number of classifiers that need to be kept, just as in variant B the set of properties is smaller than the set of words that are mapped into it, but here the clusters are chosen based on the data, and not on prior assumptions. The object description then is represented as a sequence of the labels of those clusters that the words in the description map into.

Table 1 shows an overview of the variants; their input and representation which make up how the descriptions are compressed and stored, and application and visual input which comprises how the scenes are perceived and applied.

3.4.3 Learning Perceptual Groundings

Variant A For variant A, we learned grounded word (stem) meanings in a similar way as done in Kennington et al. (2015): For each word stem w occurring in the TAKE corpus of referring expressions, we train a binary logistic regression classifier (see (5) below, where $\mathbf{w}$ is the weight vector that is learned and σ is the logistic function) that takes a visual feature representation of a candidate object ($\mathbf{x}$) and is asked to return a probability p_w for this object being a good fit to the word. We present the object that the utterance referred to as a positive training example for a good fit, and objects that it didn’t refer to as a negative example. (See Kennington et al. (2015) for a discussion of the merits of this strategy.)

$$p_w(\mathbf{x}) = \sigma(\mathbf{w}^\intercal \mathbf{x} + b) \tag{5}$$

As mentioned above, each classifier is fully specified by its coefficients ($\mathbf{w}$ and b).

Variant B As described above, the first step in variant B was to use the words in the object description as evidence for how to fill the semantic

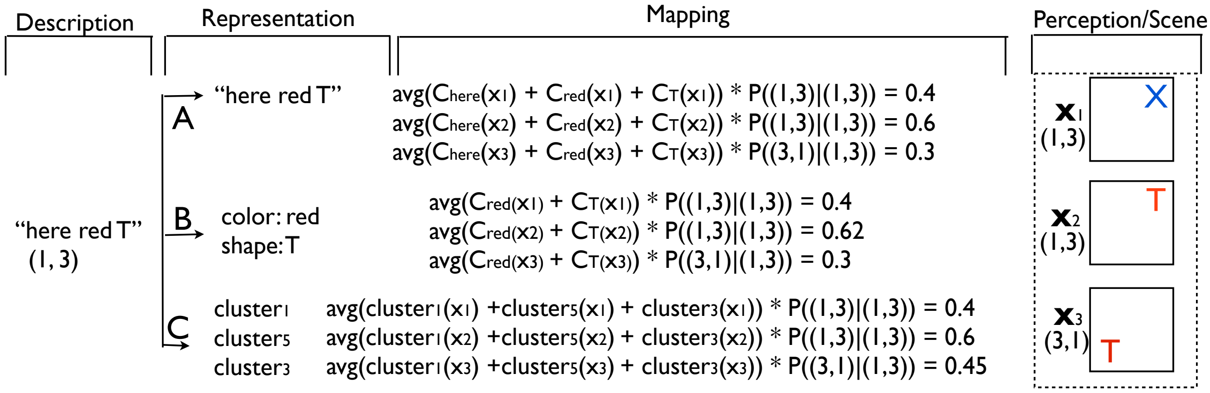

Description
Representation
Mapping
Perception/Scene
"here red T"
(1, 3)
A
"here red T"
avg(Chere(x1) + Cred(x1) + CT(x1)) * P((1,3)|(1,3)) = 0.4
avg(Chere(x2) + Cred(x2) + CT(x2)) * P((1,3)|(1,3)) = 0.6
avg(Chere(x3) + Cred(x3) + CT(x3)) * P((3,1)|(1,3)) = 0.3
B
color: red
shape: T
avg(Cred(x1) + CT(x1)) * P((1,3)|(1,3)) = 0.4
avg(Cred(x2) + CT(x2)) * P((1,3)|(1,3)) = 0.62
avg(Cred(x3) + CT(x3)) * P((3,1)|(1,3)) = 0.3
C
cluster1
cluster5
cluster3
avg(cluster1(x1) +cluster5(x1) + cluster3(x1)) * P((1,3)|(1,3)) = 0.4
avg(cluster1(x2) +cluster5(x2) + cluster3(x2)) * P((1,3)|(1,3)) = 0.6
avg(cluster1(x3) +cluster5(x3) + cluster3(x3)) * P((3,1)|(1,3)) = 0.45
x1
(1,3)
X
x2
(1,3)
T
x3
(3,1)
T

Figure 6: Simplified (and constructed) pipeline example. The description "here a red T" with gesture at point (1,3) is represented and mapped to the perceived scenes. Each variant assigns a higher probability to the correct scene, represented by X_2

frame, with the frame elements *colour* and *shape*. For the possible values of these elements (e.g., `green`) we trained the same type of logistic regression classifier, again using cases where the property was present for a given object as positive example and, as negative examples, those where it wasn't. This then gives a perceptual grounding for the property `green` (whereas in variant A we trained one for the *word* "green"). In a way, this variant begs the question of where the ontology of properties comes from; if this part is a model of language acquistion, the claim would be that there is a set of innate labels which just need to be instantiated.

Variant C As described above, this variant builds on variant A, by reducing the set of classifiers that are required through clustering. In the experiment described below, we set the number of clusters to compute to 26 (an experimentally determined optimum), which resulted for example in one cluster grouping together "violett" and "lila" (*violet* and *purple*), or another one clustering "türkis, blau, dunkelblau" (*turquoise, blue, dark blue*), but also clusters that are less readily interpreted such as "nochmal, rosa, hmm" (*again, pink, erm*). What is important to note here in any case is that the reduction in the range of what words can map into in their semantic representation is as strong as with B, but emerges from the data.

3.4.4 Applying the Information

With all this in hand, the final score for a given candidate scene is computed as follows. For each possible mapping of description object IDs into computer vision object IDs, a gestural score is computed as described in Section 3.3; the representation of each description is applied to its corresponding object using the grounding just explained; this is combined into an average description score, which is weighted by the gesture score to yield the final score of this mapping for this candidate scene.

Figure 6 shows a simple example (constructed using a simplified coordinate system for the gesture) of how each variant would process a description of a single object. Each variant is applied to the three candidate scenes.

4 Experiment

4.1 A Corpus of Scene Descriptions

To elicit natural language descriptions, we generated 25 pentomino scenes as described above and illustrated in Figure 2. We asked two student assistants (native German speakers; not authors of the paper) to write down verbal descriptions of the scenes, following a specific template (*here there is DESCR, and RELATION is...*). With these data, we do not need to run the full pipeline as described above but rather simulate the output of ASR (to focus on the core of the model for the purposes of this paper).

Example (3) shows a sample description, in which $|_{NS}$ indicates the start of a scene description and $|_{NObj}$ indicates the start of an object description. In total, we collected 50 scene descriptions.

(3) a. $|_{NS}$ Hier ist $|_{NObj}$ ein pinkes z-ähnliches Zeichen und schräg rechts unten davon ist $|_{NObj}$ ein zweites pinkes z-ähnliches Zeichen und schräg rechts unten davon ist $|_{NObj}$ ein blaues L

b. $|_{NS}$ here is $|_{NObj}$ a pink Z and diagonally to the bottom right of it is $|_{NObj}$ a second pink Z and diagonally bottom right of it is $|_{NObj}$ a blue L

We also simulate the outcome of the gesture recognition module, by taking the actual positions of the described objects as gesture positions and then adding (normally distributed) noise to simulate sensor uncertainty. The likelihood model for mapping scores is learned by producing a large number of noisy "gesture positions" based on real positions (by adding 2D gaussian noise, $\mu = 0, \sigma = 0.1$), scoring these, and then running a kernel density estimation to learn which deviation scores given the true positions are more likely.

The modules that are simulated in the evaluation are grayed out in Figure 4. Again, this is done to focus on testing the representation variants; swapping in the actual ASR and gesture modules, which we do have separately but not yet integrated, will hopefully result only in quantitatively but not qualitatively different performance.

4.2 Evaluation

To evaluate the performance of the model and the variants A-C, we created a set of test scenes for each description (hence, resulting in 50 test retrieval tasks), in three variants (for Experiments 1–3 below). Each test set includes the scene that was actually described plus as distractors five other scenes randomly selected from the set of 25 scenes (Experiment 3). For Experiment 1, the distractor scenes are modified so that all objects have the same position; i.e., in these cases, gesture information cannot help make a distinction and all load is on the verbal content. For Experiment 2, the object positions are kept, but all objects are replaced to be identical to those from the intended scene; i.e., here verbal content cannot distinguish between scenes. We created these different sets to be able to evaluate the relative contributions of each modality (Experiments 1 & 2) as well as the joint performance (Experiment 3).

We run the pipeline on the description to build the representation (or rather, three different representations, according to variants A-C) and to use this representation to retrieve the described scene from the set of six candidate scenes. We give results below in terms of accuracy (ratio of correct retrievals) as well as mean reciprocal rank (MRR), which is computed as follows (and ranges in our case from $1/6$ (worst) to 1 (ideal):

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank(i)} \qquad (6)$$

4.3 Results

Table 2 shows the results of the experiments. Experiment 1 shows that when only language contributes to the retrieval task, the representation variants can already achieve good performance, with Variant A (verbatim representation / word classifiers) having a slight edge on Variant C (representation through clustering). Going just with the gesture information, by design, performs on chance level here (top row). Experiment 2 shows that all three variants perform robustly only with gesture information: In many sets, gesture information alone already identifies the correct scene (top row, "gesture"). Language can improve over this in cases where gesture-alone computes the wrong mapping of description IDs and object IDs. Experiment 3 finally shows application to the unchanged test set with randomly selected distractors. Here, verbal information can contribute even more, and variants A and C show a perfomance that is much better than variant B. (Variant B suffers from data sparsity: e.g., in the training data, the shape U was often described as "C", and rarely as "U", which is the preferred description in our test data, leading to the wrong shape property being predicted.) Interestingly, "compressing" the information into a small number (here: 26) of clusters does not seem to have hurt the performance.

5 Related Work

As noted by Roy and Reiter (2005), language is never used in isolation; the meanings of words are learned based on how they are used in contexts–for our purposes here, *visual* contexts–where visually-perceiveable scenes are described (albeit scenes that are later visually perceived). This approach to semantics is known as *grounding*; work has been done by, inter alia, Gorniak and Roy (2004), Gorniak and Roy (2005), Reckman et al. (2010) where word meanings such as colour, shape, and spatial terms were learned by resolving referring expressions. Symbolic approaches to semantic meaning (e.g., first-order logic) do not model such perceptual word meanings well (Harnad, 1990; Steels and Kaplan, 1999); here we follow Harnad (1990) and Larsson (2013) and try to reconcile grounded semantics and symbolic approaches. In this pa-

		Experiment 1		Experiment 2		Experiment 3	
		ACC	MRR	ACC	MRR	ACC	MRR
Gesture		0.1	0.37	0.65	0.75	0.67	0.78
Gesture+Speech	A	**0.82**	**0.90**	0.70	0.78	0.84	0.91
	B	0.68	0.81	0.68	0.76	0.68	0.81
	C	0.80	0.89	**0.76**	**0.82**	**0.84**	**0.92**

Table 2: Results of the Experiments. Exp. 1: objects in same spatial configuration in all scenes (per retrieval task); Exp. 2: objects potentially in different configurations in scenes, but same three objects in all scenes; Exp. 3: potentially different objects and different locations in all scenes.

per, we extended earlier work in this area (Larsson, 2013; Kennington et al., 2015) by learning and applying these mappings in a navigation task.

Navigation tasks provide a natural environment for the development and application of such a model of grounded semantics, which have been the subject of a fair amount of recent research: In Levit and Roy (2007), later extended in Kollar et al. (2010), the meaning of words related to map-navigation such as "toward" and "between" were learned from interaction data. Vogel and Jurafsky (2010) applied reinforcement learning to the task of learning the mapping between words in direction descriptions and routes. Also, Artzi and Zettlemoyer (2013) learned a semantic abstraction from the interaction map-task data in the form of a combinatory categorical grammar. Though interesting in their own right, these tasks made some important simplifying assumptions that we go beyond in this paper: first, gestural information is never used to convey scene descriptions; second, the scene that is being described (from a bird's-eye view; here, scenes are perceived from a first-person perspective) is visually-present at the time the descriptions are being made; third, that the grounded semantics of a select subset of words are being learned. In this paper, gestures are considered, a description is heard and *later* applied to scenes, and all the word groundings are learned from data.

The work presented in this paper is a natural next step that goes beyond map-task navigation and is pshycholinguistically motivated. Kintsch and van Dijk (1978) suggest that readers (listeners) first represent exact words of a description (i.e., surface form), then interpret information (i.e., a *gist* of the description) and integrate that with their world knowledge (e.g., the knowledge about what red things look like, if the word "red" was used in the description). Moreover, Brunyé and Taylor (2008) (as well as some work cited there) note that readers construct cohesive mental models of what a text describes, integrating time, space, causality, intention, and person- and object-related information. That is, readers progress beyond the text itself to represent the described situation; detailed information from an instruction or description is distorted in memory (Moar and Bower, 1983). In this paper, we have shown in our evaluation that this is indeed the case; in Experiment 3, Variant C held the description in a more compact form than a (stemmed) surface form and produced better scores than Variant A, which did.

6 Conclusions

We have presented a first attempt at providing an end-to-end model of the task of understanding verbal/gestural scene descriptions, where this understanding can be tested by application of the understanding in a real-world (visual) discrimination task. We have explored different ways of representing content, where we went from not compressing the description at all (storing sequences of (stemmed) words, as they occurred), over using pre-specified property symbols to learning a set of "concepts" automatically. The approach overall performed well, with gesture information providing a large amount of information, with verbal content in all variants further improving over that.

In future work, we will test if the performance of the clustering approach can be improved by providing a larger amount of training data. We will also integrate the steps that were simulated here (speech and gesture recognition), and will integrate the processing pipeline into an interactive system that can potentially clarify the scene description it receives, while building the representation and before having to apply it.

Acknowledgment This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly Supervised Learning of Semantic Parsers for Mapping Instructions to Actions. *Transactions of the ACL*, 1:49–62.

Tad T Brunyé and Holly A Taylor. 2008. Working memory in developing and applying mental models from spatial descriptions. *Journal of Memory and Language*, 58(3):701–729.

Charles Fillmore. 1982. Frame semantics. *Linguistics in the morning calm*, pages 111–137.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Peter Gorniak and Deb Roy. 2005. Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. In *In Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*, pages 138–143.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D*, 42:335–346.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *SIGdial 2013*.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of IWCS*. Association for Computational Linguistics.

Walter Kintsch and Teun a. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. *Proceeding of the 5th ACMIEEE international conference on Humanrobot interaction HRI 10*, page 259.

Spyridon Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. 2012. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*, pages 39–42.

Spyros Kousidis, Thies Pfeiffer, and David Schlangen. 2013. MINT . tools : Tools and Adaptors Supporting Acquisition , Annotation and Analysis of Multimodal Corpora. In *Proceedings of Interspeech 2013*, pages 2649–2653, Lyon, France. ISCA.

S Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Michael Levit and Deb Roy. 2007. Interpretation of spatial language in a map navigation task. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(3):667–679.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.

Ian Moar and Gordon H Bower. 1983. Inconsistency in spatial knowledge. *Memory & Cognition*, 11(2):107–113.

Hilke Reckman, Jeff Orkin, and Deb Roy. 2010. Learning meanings of words and constructions, grounded in a virtual game. *Semantic Approaches in Natural Language Processing*, page 67.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Laura F Schneider and Holly a. Taylor. 1999. How do you get there from here? Mental representations of route descriptions. *Applied Cognitive Psychology*, 13(September 1998):415–441.

Luc Steels and Frederic Kaplan. 1999. Situated grounded word semantics. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2, pages 862–867.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Adam Vogel and Dan Jurafsky. 2010. Learning to Follow Navigational Directions. In *Proceedings of ACL*, pages 806–814.

User Information Extraction for Personalized Dialogue Systems

Toru Hirano, Nozomi Kobayashi, Ryuichiro Higashinaka,
Toshiro Makino and Yoshihiro Matsuo
NTT Media Intelligence Laboratories, Nippon Telegraph and Telephone Corporation
1-1 Hikarinooka, Yokosuka-Shi, Kanagawa, 239-0847, Japan
{`hirano.tohru`, `kobayashi.nozomi`, `higashinaka.ryuichiro`
`makino.toshiro`, `matsuo.yoshihiro`}`@lab.ntt.co.jp`

Abstract

We propose a method to extract user information in a structured form for personalized dialogue systems. Assuming that user information can be represented as a quadruple ⟨predicate-argument structure, entity, attribute category, topic⟩, we focus on solving problems in extracting predicate argument structures from question-answer pairs in which arguments and predicates are frequently omitted, and in estimating attribute categories related to user behavior that a method using only context words cannot distinguish. Experimental results show that the proposed method significantly outperformed baseline methods and was able to extract user information with 86.4% precision and 57.6% recall.

1 Introduction

Recent research on dialogue agents has focused extensively on casual conversations or chat (Ritter et al., 2011; Wong et al., 2012; Meguro et al., 2014; Higashinaka et al., 2014) because chat-oriented conversational agents are useful for entertainment or counseling purposes. To make users want to talk to such conversational agents more, users and systems need to know each other well since it is important to build relationships of trust between users and systems (Bickmore and Picard, 2005).

In casual conversations between people, people sometimes talk about themselves such as mentioning their hobbies or experiences. Our manual examination of text-based casual conversation between two people indicated that 26% of utterances are self-disclosure utterances that convey information about the speaker. We also observed the same tendency in casual conversations between a person and a system.

On the basis of these findings, we propose a method to extract information about the speaker, that is, user information, from utterances in order to develop personalized dialogue systems using the extracted user information. For instance, for a user who said “I went to London.”, “I live in Tokyo.”, and “I love One Direction.” (a pop band), we want to personalize conversations as follows.

Ex. 1 Telling users that the system remembers conversations in the past.

USER**:** I would like to go traveling!
SYSTEM**:** You **went to London** the other day, didn't you?

Ex. 2 Complementing unknown conditions with user information.

USER : What time does Frozen start?
SYSTEM : At 10 AM at **Tokyo Theater**.

Ex. 3 Providing information related to user interests.

USER : It's time to practice karaoke.
SYSTEM : A new song by **One Direction** is coming out soon.

To implement personalized dialogue systems such as this, the extracted user information should satisfy the following requirements.

1. It should have information to reproduce what users said in order to tell users their past utterances.
2. It should have information to complement unknown conditions.
3. It should have information that can be searched to determine which information to provide.
4. It should have information to determine when systems will use which user information.

To satisfy requirement 1, we extract the predicate-argument structure (PAS), which represents “who did what to whom.” PAS is useful for representing the basic content of an utterance. Higashinaka et al. (2014) proposed a method to generate system utterances from PAS. For requirements 2 and 3, we extract entities such as person

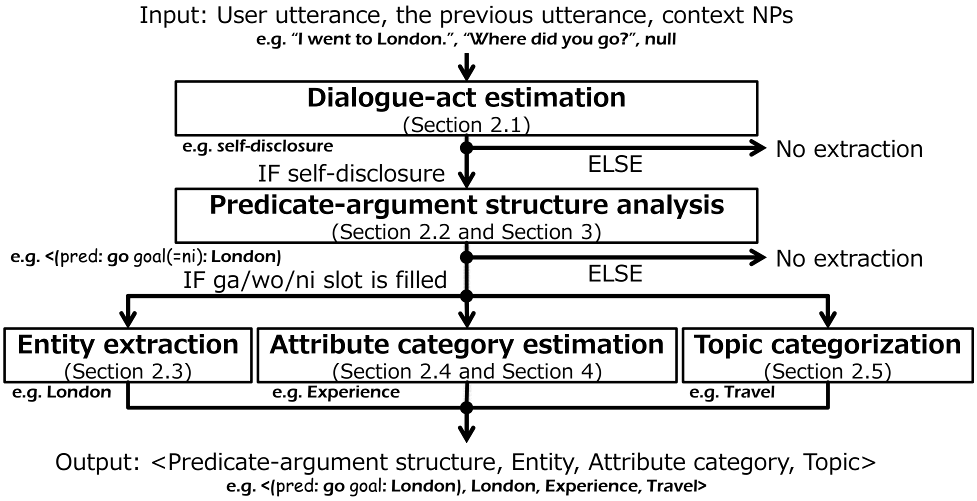

Input: User utterance, the previous utterance, context NPs
e.g. "I went to London.", "Where did you go?", null
Dialogue-act estimation
(Section 2.1)
e.g. self-disclosure
IF self-disclosure
ELSE
No extraction
Predicate-argument structure analysis
(Section 2.2 and Section 3)
e.g. <(pred: go goal(=ni): London)
IF ga/wo/ni slot is filled
ELSE
No extraction
Entity extraction
(Section 2.3)
Attribute category estimation
(Section 2.4 and Section 4)
Topic categorization
(Section 2.5)
e.g. London
e.g. Experience
e.g. Travel
Output: <Predicate-argument structure, Entity, Attribute category, Topic>
e.g. <(pred: go goal: London), London, Experience, Travel>

Figure 1: Overview of user information extraction.

and location names, which represent keywords. For requirement 4, we extract attribute categories such as hobbies and experiences, which represent aspects of users, to determine which information will be used, and we extract topics such as music and travel, which represent main subjects, to determine when to use the information.

Therefore, we extract a quadruple ⟨PAS, entity, attribute category, topic⟩ as user information from a user utterance. From the above examples, we extract the following information.

- "I went to London." → ⟨(pred: go goal: London), London, Experiences, Travel⟩
- "I live in Tokyo." → ⟨(pred: live locative: Tokyo), Tokyo, Place of residence, House⟩
- "I love One Direction." → ⟨(pred: love accusative: One Direction), One Direction, Hobbies/Preferences, Music⟩

In this paper, the work is done in Japanese although we want to apply our method to other languages in the future. For languages other than Japanese, instead of PASs, semantic role labeling (SRL) can be used (Palmer et al., 2010).

2 User Information Extraction

An overview of the method we propose to extract user information, ⟨PAS, entity, attribute category, topic⟩, from user utterances is shown in Figure 1. The method has five parts:

- Dialogue-act estimation
- Predicate-argument structure analysis
- Entity extraction
- Attribute category estimation
- Topic categorization

We focus on solving problems in analyzing predicate argument structures of question-answer pairs in which arguments and predicates are frequently omitted, and in estimating attribute categories related to user behavior that a method using only context words cannot distinguish. In this section, we outline the overall functionality of a user information extraction system; further methods to solve the problems are described in sections 3 and 4.

2.1 Dialogue-act estimation

We identify the dialogue-act of utterances to determine whether the input user utterance contains information about the user him/herself. We use the dialogue-act tag set consisting of 33dialogue-acts listed in Table 1, proposed by Meguro et al. (2014). Their tag set is designed for annotating listening-oriented dialogue, but because speakers in listening-oriented dialogue are allowed to speak freely, the tag set can cover diverse utterances, making it suitable for casual conversation.

We evaluate whether the input user utterance contains information about the user in two cases, as follows.

1. the dialogue-act of the user utterance is one of the self-disclosure tags: No. 3–11.
2. the dialogue-act of the user utterance is one of the sympathy/agreement tags: No. 22–23, and the dialogue-act of the previous utterance is one of the question tags: No. 14–19.

No.	Dialogue-acts
1	greeting
2	information
3	self-disclosure.fact
4	self-disclosure.experience
5	self-disclosure.habit
6	self-disclosure.preference.positive
7	self-disclosure.preference.negative
8	self-disclosure.preference.neutral
9	self-disclosure.desire
10	self-disclosure.plan
11	self-disclosure.other
12	acknowledgment
13	question.information
14	question.fact
15	question.experience
16	question.habit
17	question.preference
18	question.desire
19	question.plan
20	question.self-questioning
21	question.other
22	sympathy/agreement
23	non-sympathy/non-agreement
24	confirmation
25	proposal
26	repeat
27	paraphrase
28	approval
29	thanks
30	apology
31	filler
32	admiration
33	other

Table 1: Dialogue-act tag set.

We use a method proposed by Higashinaka et al. (2014) to estimate a dialogue-act. They trained a classifier using a support vector machine (SVM). The features used are word N-grams, semantic categories obtained from the Japanese thesaurus Goi-Taikei (Ikehara et al., 1999), and character N-grams.

2.2 Predicate-argument structure analysis

Predicate-argument structure (PAS) analysis involves detecting predicates and their arguments. A predicate can be a verb, adjective, or copular verb, and the arguments are noun phrases (NPs) associated with cases in case grammar. As cases, we use ga (nominative), wo (accusative), ni (dative), de (locative/instrumental), to (with), kara (source), and made (goal).

We use the PAS analyzer described by Imamura et al. (2014) to analyze PASs for general utterances. The analyzer works statistically by ranking NPs in the context using supervised learning with an obligatory case information dictionary and a large-scale word dependency language model. On the other hand, the analyzer cannot extract PASs correctly in order to analyze them for question-answer pairs in which arguments and predicates are frequently omitted. For example, predicate ellipsis is not targeted by the analyzer. Therefore, we use a method described in section 3 to analyze the PAS for question-answer pairs.

No.	Attributes	No.	Attributes
1	relationship:family	18	occupation
2	relationship:partner	19	place of business
3	relationship:lover	20	position in company
4	relationship:other	21	journey to work
5	name	22	biography
6	gender	23	earnings
7	age	24	expenditure
8	blood type	25	possessions
9	birthday	26	knowledge
10	constellation	27	hobbies/preferences
11	Chinese zodiac	28	habits
12	characters	29	experiences
13	physical description	30	strong points
14	home town	31	abilities
15	place of residence	32	opinions/feelings
16	house mate	33	desires
17	house type	34	other

Table 2: Attribute category tag set.

To extract user information, we need to select a PAS from the ones in the input utterances. We select the last PAS in the utterance on the basis of the observation that important information comes last in many Japanese utterances. Additionally, we should not output insufficient PASs in which argument slots are not filled at all because the extracted PASs would be used to generate system utterances. Therefore, we output PASs only when at least one of the argument slots (ga, wo, or ni) of the predicate is filled.

2.3 Entity extraction

We define an entity as a noun phrase (NP) that denotes the center word of a conversation. To extract the entity from the input user utterance, we use the center word extraction method proposed by Higashinaka et al. (2014). They extracted an NP from an utterance and trained a conditional random field (Lafferty et al., 2001); NPs are extracted directly from a sequence of words without creating a parse tree. The feature template uses words, part-of-speech (POS) tags, and semantic categories of current and neighboring words.

When no NP is extracted from the input user utterance, we try extracting an NP from previous utterances.

2.4 Attribute category estimation

We identify an attribute category, which represents aspects of users, for self-disclosure utterances of the user, e.g. “I went to London.” → experi-

No.	Topics	No.	Topics
1	travel	23	disaster prevention
2	events	24	volunteering
3	movies	25	health
4	music	26	post-retirement
5	TV	27	beauty
6	entertainment	28	fashion
7	talent	29	shopping
8	computers	30	gourmet dining
9	games	31	anime
10	telephone	32	occult
11	business	33	gardening
12	study	34	sports
13	school	35	art
14	money	36	books
15	animals	37	cars/bikes
16	home	38	history
17	housekeeping	39	fishing
18	appliances	40	fortune-telling
19	family	41	religion
20	friends	42	general
21	love	43	other
22	politics		

Table 3: Topic category tag set.

ences. As an attribute category set, we define 34 categories of attributes in Table 2 on the basis of a questionnaire conducted in a market research study and on the analysis of personal questions (Sugiyama et al., 2014). The inter-annotator agreement with 200 self-disclosure utterances was 90.5% (Cohen's κ = 0.885). Because κ is more than 0.8, we can say the agreement is high.

We used a logistic-regression-based classifier to estimate attribute categories. We describe in section 4 the features used to estimate attribute categories related to user behavior that a method using only context words cannot distinguish.

2.5 Topic categorization

We identify a topic category, which represents the main subject, of the input user utterances, e.g. "I went to London." → travel. As a topic category tag set, we use 43 categories listed in Table 3 based on categories used on a Japanese question and answer communication site[1]. The inter-annotator agreement with 200 utterances was 93.0% (Cohen's κ = 0.925). Because κ is more than 0.8, we can say the agreement is high.

To categorize topics, we trained a classifier in the same way as done with attribute category estimation.

3 Analyzing Predicate-argument Structure of Question-answer Pairs

As mentioned in section 2.2, analyzing PASs for question-answer pairs in which predicates and arguments are frequently omitted is problematic. Although many prior studies have been done on PAS analysis (Taira et al., 2010; Hayashibe et al., 2011; Yoshikawa et al., 2011; Imamura et al., 2014), the methods they use could not be applied to analyze PASs of question-answer pairs with ease. For example, they could not extract the following PAS because of predicate ellipsis. The PAS (pred: read accusative: Fashion magazines) should be extracted from the example.

> "Do you read books?" - "Fashion magazines." → (pred: ϕ)

Types	Rate
completed	47.1% (115/244)
argument ellipsis	28.3% (69/244)
predicate ellipsis	7.8% (19/244)
yes-no	16.8% (41/244)

Table 4: Types of question-answer pairs.

To solve the problem, we break question-answer pairs down into the following four types, and we propose a method to analyze PAS for each of them except for the completed type.

Completed: Both the predicate and its arguments are included. (e.g. "What is your hobby?" - "My hobby is playing tennis.")

Argument ellipsis: A predicate is included, but the argument is omitted. (e.g. "Did you go to London last year?" - "I went with friends.")

Predicate ellipsis: A predicate is omitted, but the argument is included. (e.g. "Do you read books?" - "Fashion magazines.")

Yes-no: An answer is either "yes" or "no". (e.g. "Do you like to read books?" - "Yes.")

Table 4 lists the percentage of these four types among 244 question-answer pairs and indicates that predicates or arguments are omitted in 52.9% (= 100% − 47.1%) of the question-answer pairs.

The question-answer pairs had some typical forms such as "What do you like?" - "(I like) x.". We can accurately extract PASs from these typical cases using predefined extraction patterns. On the basis of these extractions, we propose a four-step method to analyze the PASs of question-answer pairs.

- pattern-based extraction: all types
- argument complement: argument ellipsis
- complete sentence generation: predicate ellipsis
- question PAS copying: yes-no

---

[1] http://oshiete.goo.ne.jp/

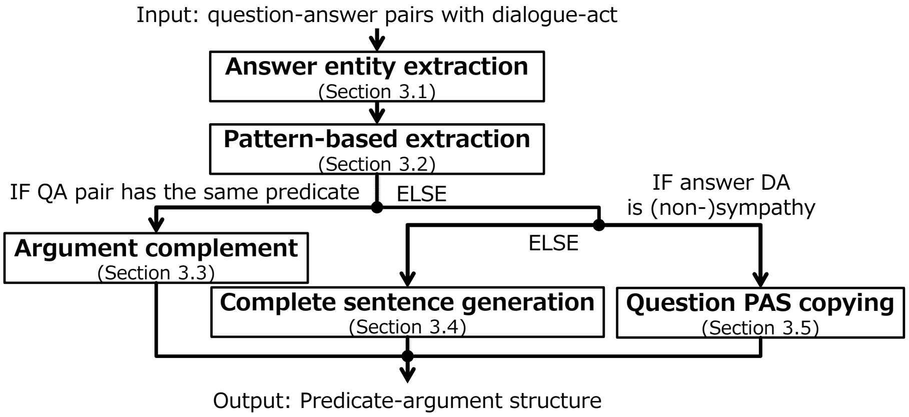

Input: question-answer pairs with dialogue-act
Answer entity extraction
(Section 3.1)
Pattern-based extraction
(Section 3.2)
IF QA pair has the same predicate
ELSE
IF answer DA is (non-)sympathy
ELSE
Argument complement
(Section 3.3)
Complete sentence generation
(Section 3.4)
Question PAS copying
(Section 3.5)
Output: Predicate-argument structure

Figure 2: Process of PAS analysis.

Note that the method can be used to analyze the completed type as well as other types, because it cannot determine if it is the completed type before analyzing the PAS. Figure 2 outlines the PAS analysis process.

3.1 Pre-process: answer entity extraction

As a pre-process, we extract an answer entity from an answer utterance using named entity recognition since the answer is likely to be regarded as a named entity. We use the named entity recognition method proposed by Sadamitsu et al. (2013), which is based on Sekine's Extended Named Entity Hierarchy[2].

3.2 Pattern-based extraction method

In the pattern-based extraction step, an attempt is made to extract predicate-argument structures using pre-defined extraction patterns. If a pattern can extract a PAS, the extracted PAS is output as an answer.

We collected frequently appearing patterns in the Person-Database (Sugiyama et al., 2014) using the frequent-pattern mining method (Pei et al., 2001) and assembled 20 regular expression patterns by checking the collected frequent patterns. The Person-Database consists of a number of question-answer pairs created by 42 questioners and includes 26,595 question-answer pairs, which cover most of the questions related to the information about users.

The following is an example of a regular expression pattern.

> “What .* do you like?” - “$answer\ entity$”
> $\rightarrow$ (pred: like accusative: $answer\ entity$)

Here, $answer\ entity$ denotes an answer entity detected in the answer entity extraction.

We show the example, “What kind of food do you like?” - “Sushi.” “Sushi” in the answer utterance is extracted as an answer entity. Therefore, (pred: like accusative: Sushi) is extracted as a PAS from this example.

When the pattern-based method extracts a predicate-argument structure, the steps described in the following subsections would be skipped.

3.3 Argument complement method

If an answer utterance has the same predicate that appeared in the question utterance, the argument complement step is executed.

This step compares the question PAS and the answer PAS that were analyzed using an existing predicate-argument structure analysis method, and complements the arguments that only appear in the question PAS. For example, when the question PAS is (pred: go goal: London) and the answer PAS is (pred: go with: friends), (pred: go goal: London with: friends) is generated by copying “goal: London” from the question PAS.

3.4 Complete sentence generation method

If an answer utterance does not have the same predicate that appeared in the question utterance and the dialogue act of the answer utterance is not “(non-)sympathy/agreement”, the complete sentence generation step is executed.

When there is a predicate-ellipsis example, we generate a complete sentence by replacing a question expression with an answer entity. A ques-

---

[2] https://sites.google.com/site/extendednamedentityhierarchy/

tion expression consists of a question word (such as "what" or "how") and suffixes or nouns (such as "food" or "meter"). For example, given the question-answer pair "What kind of food do you like?" - "Sushi", "(I) like Sushi." is generated as a complete sentence by replacing the question expression "What kind of food" with the answer entity "Sushi" and then converting a question sentence into an affirmative sentence.

The question expression is extracted with a predefined question word list and extraction rules. The rules extract suffixes or nouns attached to a question expression as a question expression. We can obtain a PAS applying existing predicate-argument structure analysis methods to the generated utterance. From the above example, we can obtain the PAS (pred: like accusative: Sushi).

3.5 Question PAS copying method

If an answer utterance does not have the same predicate that appeared in the question utterance and the dialogue act of the answer utterance is "(non-)sympathy/agreement", the question PAS copying step is executed.

A yes-no type answer PAS is empty because the answer utterance is expressed by an interjection such as "yes" or "no". This case is regarded as a case in which a predicate and its arguments are both omitted, so the question PAS is output as the answer PAS. For the example, "Did you go to London?"- "Yes.", the question PAS (pred: go goal: London) is extracted as the answer PAS.

4 Estimating Attribute Categories Related to User Behavior

Attribute category estimation is used to identify an attribute category for self-disclosure utterances of the user. For example, the utterance "I went to London." should be categorized with an experiences tag. A simple approach to estimate the attribute category for self-disclosure utterances of the user is a logistic-regression-based classifier with word N-gram features and semantic category features obtained from the Japanese thesaurus Goi-Taikei (Ikehara et al., 1999), which are used for topic categorization. These context features are important clues for identifying 26 categories, No. 1–26 in Table 2, but they are not important clues for identifying the other categories, No. 27–34, which are related to user behavior.

For instance, the baseline method incorrectly classifies the utterance "I played tennis a little while ago.", which should be classified with an experience tag, and the utterance "I always play tennis.", which should be classified with a habit tag, because the context words in both utterances are the same, "play" and "tennis".

To solve this problem, we need to use features representing whether the user behavior has ended, is continuing, or was repeated. Therefore, we propose using semantic information of functional words and adverbs as features to classify attribute categories related to user behavior.

4.1 Semantic information of functional words

We use semantic information of functional words in our proposed method. In Japanese, "-ta" is a past tense expression that means the action was completed, and "-teiru" is a present tense expression that means the action is continuing, so semantic information of functional words would be important clues to classify attribute categories related to user behavior. In this paper, we use semantic labels of function words by analyzing function words using the method proposed by Imamura et al. (2011) as features. We assume that semantic labels of functional words would be important clues to classify attribute category, especially the semantic labels "completion" for the experiences category, "continuance" for the habits category, "supposition" and "admiration" for the opinions/feelings category, and "request" and "desire" for the desires category.

4.2 Semantic information of adverbs

We use semantic information of adverbs such as "a little while ago" or "always" in our proposed method. For instance, in our method, "always" expresses that the action is done on a daily basis, and "a little while ago" expresses time information about when the action was done. In attribute category estimation, we expect that adverbs expressing that the action is done on a daily basis would be important clues for the habits category, and adverbs expressing the time in which the action was done would be important clues for the experiences category.

We prepare in advance two lists of adverbs that are used in order to extract semantic information of adverbs: (A) a list of adverbs expressing that the action is done on a daily basis, e.g. "always" and "every day", and (B) a list of adverbs expressing the time the action was done, e.g. "a little while ago" and "before". Such lists represent the semantic information of adverbs, so we use the lists of extracted adverbs as features.

	Baseline			Proposed		
Types	Precision	Recall	F	Precision	Recall	F
completed	**89.2**% (149/167)	**87.6**% (149/170)	**0.884**	85.3% (139/163)	81.8% (139/170)	0.835
argument ellipsis	43.1% (66/153)	40.7% (66/162)	0.419	**63.1**% (94/149)	**58.0**% (94/162)	**0.605**
predicate ellipsis	26.3% (5/19)	19.2% (5/26)	0.222	**41.7**% (10/24)	**38.5**% (10/26)	**0.400**
yes-no	40.0% (30/75)	28.0% (30/107)	0.330	**67.0**% (61/91)	**57.0**% (61/107)	**0.616**
total	60.4% (250/414)	53.8% (250/465)	0.569	**71.2**% (304/427)	**65.4**% (304/465)	**0.682**

Table 5: Comparison of PAS analysis of question-answer pairs for baseline and proposed methods.

5 Experiments

5.1 Predicate-argument structure analysis of question-answer pairs

We investigated how effective the proposed method described in section 3 was in analyzing PAS of question-answer pairs by comparing it with a baseline method. The baseline method used was that of Imamura et al. (2014), which is described in section 2.2. This method analyzes the PASs of question-answer pairs as well as the other utterances.

We used 478 question-answer pairs in 480 casual dialogues between a person and a system (Higashinaka et al., 2014) to evaluate whether the system could extract PASs correctly.

Table 5 lists the performance results of both methods for various types of question-answer pairs. Precision is defined as the percentage of correct PASs out of the extracted ones. Recall is the percentage of correct PASs from among the manually extracted ones. The F measure is the harmonic mean of precision and recall.

A comparison between the baseline and proposed methods indicates that the F measure of the proposed method improved by 0.113 points. The use of a statistical test (McNemar Test) demonstrably showed the proposed method's effectiveness. Specifically, the proposed method increased the F measure by 0.186 points in the argument ellipsis type and by 0.286 points in the yes-no type.

Our error analysis indicated that 40% of errors consisted of a failure to include complementing arguments of predicates. For examples, from the pair "Did you have dinner tonight?" - "I ate a little while ago." the system would not extract the correct PAS because the predicate is not the same in the question (have) and the answer (eat). To solve this problem, we plan to evaluate whether two predicate-argument structures have the same meaning by applying paraphrase detection methods such as using recursive autoencoders (Socher et al., 2011). In addition, we plan to improve the handling of ellipsis and anaphora by incorporating methods that utilize syntactic structures (Dalrymple et al., 1991; Iida et al., 2007).

Method	Accuracy
baseline	76.0% (14,120/18,579)
proposed	**88.9**% (16,523/18,579)
upper bound (ref.)	90.5% (181/200)

Table 6: Accuracy of attribute category estimation.

5.2 Attribute category estimation

We also investigated the effectiveness of the proposed method described in section 4 when using semantic information of functional words and adverbs as features by comparing its results with those of the baseline method. To train a logistic-regression-based classifier, we used LIBLINEAR[3] with both methods.

We used 18,579 self-disclosure utterances as well as previous utterances from 4,160 casual dialogues: 3,680 dialogues between two people and 480 dialogues between a person and a system, annotated with 34 categories listed in Table 2. The number of utterances annotated for each category in decreasing order was: opinions/feelings 5,580 (30%); experiences 4,758 (25%); habits 2,414 (12%); and hobbies/preferences 2,234 (12%). We used the above self-disclosure utterances for training and testing by ten-fold cross validation.

Table 6 gives the accuracy of the baseline and proposed methods in estimating the attribute category, and the inter-annotator agreement as a referential upper bound. A comparison between the baseline and proposed methods indicates that the proposed method using semantic information of functional words and adverbs improved the accuracy by 12.9 points. The use of a statistical test (McNemar Test) demonstrably showed the proposed method's effectiveness. With the proposed method, the accuracy was greatly improved to 86.9% from 66.4% in the habits category and to 89.0% from 68.9% in the experiences category.

A comparison between the referential upper bound and the proposed method indicates that the

---

[3] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

	Precision	Recall	F
baseline	57.4% (171/298)	34.2% (171/500)	0.429
proposed	**86.4**% (288/333)	**57.6**% (288/500)	**0.691**

Table 7: Performance of user information extraction.

proposed method is very close to the upper bound accuracy.

5.3 Overall performance of user information extraction system

To evaluate the overall functionality of a method implemented with the user information extraction system described in section 2, we used 500 user utterances randomly selected from 3,680 casual dialogues (Higashinaka et al., 2014) between two people, and annotated with PAS, entity, attribute categories, and topics.

Table 7 lists the performance results of the baseline and proposed methods in extracting user information from user utterances. A comparison between the two methods indicates that the proposed method improved the F measure by 0.262 points. The use of a statistical test (McNemar Test) demonstrably showed the proposed method's effectiveness.

This result demonstrates that the proposed method was able to extract user information with high precision, 86.4%, and moderate recall, 57.6%. User information extracted with such high precision would be useful for personalized dialogue systems, because when the extracted information is wrong, the system personalizes it in the wrong way.

The proposed method could not extract user information from 167 (= 500 − 333) utterances because of incorrect dialogue-act estimation (18 utterances) and PAS analysis (149 utterances). We need to solve these problems, especially in the PAS analysis, to extract more user information.

6 Related Work

Several studies have been done on extracting user information from user utterances (Weizenbaum, 1966; Wallace, 2004; Kim et al., 2014; Corbin et al., 2015). Chat bot systems such as ELIZA (Weizenbaum, 1966) and ALICE (Wallace, 2004) extract the user information, name, and hobby of the user by using predefined pattern rules to personalize casual dialogues. In these systems, since the extracted information is limited to match predefined pattern rules, new rules need to be added in order to extract new information.

In a dialogue system used to find out information on the colleagues of the user (Corbin et al., 2015), the system extracts where the user sits in an office and uses the extracted information to search a database for a personalized information service. In this study, the same problem exists in that the extracted user information is limited to where the user sits.

Kim et al. (2014) used open information extraction (OpenIE) techniques (Banko and Etzioni, 2008) to solve this problem. OpenIE extracts triples ⟨NP, relation, NP⟩ that include relation expressions between NPs, without using predefined pattern rules. Using this framework, the system was able to extract ⟨I, like, apples⟩ in a structured form from the utterance "I like apples." in order to generate system utterances. In this study, because their purpose is only to generate utterances directly from extracted user information, they do not extract attribute categories and topics. Thus, it can be said that our work expands the types of personalized conversation by extracting quadruples ⟨PAS, entity, attribute category, topic⟩.

Much research has been done on information search (Shen et al., 2005; Qiu and Cho, 2006) and recommendation (Ardissono et al., 2004; Jiang et al., 2011) in the research area of personalization. These studies represent user interests with word vectors by comparing a vector of user interests and document vectors and selecting a document that has a similar vector to a user interest vector. These methods can roughly capture user interests, but they cannot precisely capture user information.

7 Conclusion

This paper proposed a method to extract user information in a structured form, ⟨predicate-argument structure, entity, attribute category, topic⟩, for personalized dialogue systems. We focused in particular on the tasks of extracting predicate argument structures from question-answer pairs and estimating attribute categories from self-disclosure utterances of the user. The experiments demonstrated that the proposed method outperformed a baseline method in both tasks and that the method was able to extract user information from human-human dialogue with 86.4% precision and 57.6% recall.

In future, we plan to implement a personalized dialogue system using extracted user information. We also want to solve the problems in PAS analysis to extract more user information and to apply our method to other languages.

References

Liliana Ardissono, Cristina Gena, Pietro Torasso, Fabio Bellifemine, Angelo Difino, and Barbara Negro. 2004. User modeling and recommendation techniques for personalized electronic program guides. In *Personalized Digital Television - Targeting Programs to Individual Viewers*, volume 6 of *Human - Computer Interaction Series*, pages 3–26.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics: Human Language Technologies*, pages 28–36.

Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327.

Carina Corbin, Fabrizio Morbini, and David Traum. 2015. Creating a virtual neighbor. In *Proceedings of the 2015 International Workshop Series on Spoken Dialogue Systems Technology*.

Mary Dalrymple, Stuart M. Shieber, and Fernando C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4):399–452.

Yuta Hayashibe, Mamoru Komachi, and Yuji Matsumoto. 2011. Japanese predicate argument structure analysis exploiting argument position and type. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 201–209.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open domain conversational system fully based on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 928–939.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing*, 6(4):1–22.

Satoru Ikehara, Masahiro Miyazaki, Satoru Shirai, Akio Yoko, Hiromi Nakaiwa, Kentaro Ogura, Masafumi Oyama, and Yoshihiko Hayashi. 1999. *Nihongo Goi Taikei (in Japanese)*. Iwanami Shoten.

Kenji Imamura, Tomoko Izumi, Genichiro Kikui, and Satoshi Sato. 2011. Semantic label tagging to functional expressions in predicate phrases. In *Proceedings of the 17th Annual Meeting of Association for Natural Language Processing (in Japanese)*, pages 518–521.

Kenji Imamura, Ryuichiro Higashinaka, and Tomoko Izumi. 2014. Predicate-argument structure analysis with zero-anaphora resolution for dialogue systems. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 806–815.

Yechun Jiang, Jianxun Liu, Mingdong Tang, and Xiaoqing Liu. 2011. An effective web service recommendation method based on personalized collaborative filtering. In *Proceedings of the 2011 IEEE International Conference on Web Services*, pages 211–218.

Yonghee Kim, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2014. Acquisition and use of long-term memory for personalized dialog systems. In *Proceedings of the 2014 Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. 2014. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing*, 10(4):15:1–15:20.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei chun Hsu. 2001. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224.

Feng Qiu and Junghoo Cho. 2006. Automatic identification of user interest for personalized search. In *Proceedings of the 15th International Conference on World Wide Web*, pages 727–736.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language*, pages 583–593.

Kugatsu Sadamitsu, Ryuichiro Higashinaka, Toru Hirano, and Tomoko Izumi. 2013. Knowledge extraction from text for intelligent responses. *NTT Technical Review*, 11(7):1–5.

Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 824–831.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the Advances in Neural Information Processing Systems 24*, pages 801–809.

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2014. Large-scale collection and analysis of personal question-answer pairs for conversational agents. In *Proceedings of the 14th International Conference on Intelligent Virtual Agents*, pages 420–433.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2010. Predicate argument structure analysis using transformation-based learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 162–167.

Richard S. Wallace. 2004. *The Anatomy of A.L.I.C.E. A.L.I.C.E.* Artificial Intelligence Foundation, Inc.

Joseph Weizenbaum. 1966. Eliza — a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9:36–45.

Wilson Wong, Lawrence Cavedon, John Thangarajah, and Lin Padgham. 2012. Strategies for mixed-initiative conversation management using question-answer pairs. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2821–2834.

Katsumasa Yoshikawa, Masayuki Asahara, and Yuji Matsumoto. 2011. Jointly extracting japanese predicate-argument relation with markov logic. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1125–1133.

How far can we deviate from the performative formula?

Lisa Hofmann
Heinrich-Heine-Universität Düsseldorf

1 Introduction

Austin (1979) proposes that performatives are unique in making explicit their illocutionary force. For example, in *'I hereby promise to bring beer.'*, we understand the illocutionary force of a promise to be explicit. Searle (1989) builds on that analysis by proposing an ontology of speech acts as actions, which can be performed by manifesting the intention to do so, and performative verbs, which denote speech acts and therefore can be used to manifest such an intention (such as *promise, order, thank* or *advise*). He suggests that performative sentences are composed with a performative main verb and some self-referentiality and could therefore potentially serve as performative utterances. In contrast, Eckhardt (2012) shows that self-referentiality can be understood as a property of utterances rather than sentences. Because self-reference is a necessary condition for performativity, also performativity can be understood as a property of utterances.

An event-based account of performative self-referentiality, in which the event-argument of the performative verb refers to the utterance raises the following questions:

- What restrictions does compositionality impose on the reference of verbal arguments in self-referential utterances?
- What can self-referentiality tell us about the properties and structure of performative events (or explicit speech acts)?

In section two, I give an in-depth critique of Eckhardt's account of performatives, which is an effort of a sufficient characterisation, of which I will adopt some parts and reject others. Section three is a proposal to extend Eckhardt's event-based account of self-referentiality along the lines of the consequences of that account, principles of event individuation and compositionality. In the third section, I will attempt to answer the first of the above questions: Coreference of the event denoted by performative verbs with the utterance event leads to restrictions on the reference of participant-arguments of performative verbs: they have to be anchored in context. Section four will conclude my proposal with a discussion of my claims and of their relevance for the semantics and pragmatics of communication and dialogue: eventive self-referentiality and context-anchored arguments tie performative meaning to the utterance and its context.

2 Eckhardt's account

2.1 A formal analysis of self-referential utterances

Eckhardt (2012) provides a truth-conditional analysis of performatives on the basis of Davidson (1980). His basic assumption, that verbs take an event-argument, allows for a straightforward implementation of self-referential utterances: because utterances are events, they are possible referents for the event argument of the main verb of the uttered sentence. The adverb *hereby*, which characteristically is taken by performative verbs, introduces a context-relative constant ε, referring to the utterance. It saturates the event argument of the verb and thereby induces performative self-referentiality.

(1) a. *I (hereby) [promise to bring beer]*$_{VP}$

b. $[\![\textit{promise to bring beer}]\!]^{w,c} = \lambda e\lambda x.\textsc{Promise}(x, e, \lambda w'.\textsc{Bring}(x, \textsc{beer}, w'), w^0)$

c. $[\![\textit{I}]\!]^{w,c} = \mathbf{sp}$ (speaker in context c)

d. $[\![\textit{hereby}]\!]^{w,c} = \varepsilon$ (ongoing act of information transfer in c)

e. $[\![\textit{I hereby promise to bring beer.}]\!]^{w,c} = \textsc{Promise}(\mathbf{sp}, \varepsilon, \textsc{Bring}(\mathbf{sp}, \textsc{beer}))$

This analysis of performatives with *hereby* corresponds to Searle's direct account. However, unlike Searle, Eckhard does not assume a tacit *hereby* for performatives occurring without it, but an existential closure, which leads to an indirect derivation of performativity, much like Bach (1975) proposed.

(2) $[\![\textit{I promise to bring beer.}]\!]^{w,c} = \exists e.\textsc{Promise}(\mathbf{sp}, e, \textsc{Bring}(\mathbf{sp}, \textsc{beer}))$

The self-referentiality of this existential statement comes about in context, along with its verification through instantiation of the existentially bound variable.

(3) $[\![\exists e.\textsc{Promise}(\mathbf{sp}, e, \textsc{Bring}(\mathbf{sp}, \textsc{beer}))]\!]^{M,g}$ = 1 because $[\![\exists e.\textsc{Promise}(\mathbf{sp}, e, \textsc{Bring}(\mathbf{sp}, \textsc{beer}))]\!]^{M,g(e/\varepsilon)}$ = 1

Eckhard shows that self-referentiality accounts for whether an utterance is interpreted as a performative or not in many cases, which were formerly considered problematic. Within this account, performative meaning is established through reference of the event-argument of the performative verb to the utterance event and diverse processes saturating verbal arguments give rise to different derivations of eventive self-referentiality. It follows that the performative event (or speech act) *is* the utterance event. However, there are performative utterances which are not strictly self-referring:

(4) a. $[\![$*King Karl hereby promises you a cow.* (to farmer Burns)$]\!]$ = $\textsc{Promise}(\textsc{King}, \varepsilon, \lambda w'.\textsc{Give}(\textsc{King}, \textsc{cow}, \textsc{Burns}))$

(4) can be interpreted performatively, when uttered by an official messenger or representative of the king. Taking into account Parsons's (1990) argument for the uniqueness of roles as a principle of event individuation, the interpretation of *hereby* in (4a) as referring to the utterance leads to a felicitous utterance only if the speaker is King Karl. Eckhard therefore generalises the reference of *hereby* to a more abstract communicative event[1]. Minimal communicative events take place at the level of utterances, and can be part of complex communicative events: A person A communicating on behalf of another person B towards a person C involves a larger communication between B and C. For performative utterances of sentences like (4), Eckhardt characterises the context-sensitive constant ε as referring to that complex communicative event. This analysis sheds light on many unsolved questions regarding the meaning of performativity. Deriving performativity through this special type of context anchoring explains how the same sentence can be used as a performative in some contexts and as a statement in others. As Eckhardt points out, a habitual interpretation of performative sentences leads to a reportative utterance.

(5) *(Whenever you invite me,) I promise to bring beer.*

A specific communicative event is not habitual. Therefore Eckhardt's theory predicts that (5), which is habitual, could not be a performative. This raises the more general question about what communicative events are (and are not) – a question which I attempt to elucidate in the forthcoming sections. Performative utterances are a special type of communicative event, which are explicitly realised in language as the main verb's event argument. Because the event denoted by the performative main verb and the communica-

---

[1] Eckhardt denominates these events as *ongoing acts of information transfer*. As has been pointed out by an anonymous reviewer, this notion is not compatible with a dialogue view of communication. I am going to use the more neutral term *communicative event*, which can not only be more convienently used to talk about its participants and properties, but also forgoes inherent assumptions about the nature of communication.

tion are the same event, they should share their thematic roles and some other eventive properties. Investigating the linguistic realisations of these arguments and properties may offer some insight into their internal structure.

2.2 Self-reference and self-verification

Another widely discussed characteristic of performative utterances is being self-verificational. Incoherent discourses like in (6) have often been used as a test for performativity.

(6) a. *I invite you to come to my party tonight. – # No, that's not true.*
b. *I invite you to come to my party tonight. – # Yes, correct.*

However, Eckhardt shows that self-verification and self-referentiality are not equivalent and both not sufficient for performativity. Many verbs can take *hereby* as an argument and thereby establish a self-referential reading. Not all of them are self-verificational or performative, as (7) shows.

(7) *I hereby utter a sentence with seven words.*

An utterance of (7) would be self-referential, but not self-verificational (it would be that of a wrong statement). The self-referentiality is explicitly realised through the use of *hereby* and the sentence is a statement about its utterance. Necessary truth, which is established through context-anchoring, on the other hand does not require self-reference either, nor does it necessarily lead to a performative interpretation.

(8) *I am here now.*

The self-verification of (8) comes about through the proximal deixis of its subparts, anchoring them in the speech situation. Although this sentence could not be used to make a self-referential utterance, its context-anchoring is similar to eventive self-reference of utterances as described by Eckhardt. Therefore the mechanism of self-verification in (8) seems to be similar to performative self-verificationalism. Given that self-verification is a necessary condition for performatives and that there are different ways in which a sentence could be self-verificational, a step towards a more precise characterisation of performativity could be asking the question how exactly performative self-verification comes about. If self-referentiality and self-verificationalism are both necessary for performativity, we should ask how the two interrelate in the construal of performative meaning.

2.3 A sufficient characterisation?

The existence of non-performative self-referential utterances as in (7) and the fact that performative sentences uttered jokingly do not establish their potential force lead Eckhardt to further make assumptions about pragmatic mechanisms as sufficient conditions for performative utterances:

She assumes that the speaker has to actively express their sincere intention to perform the described speech act. She proposes that performative utterances involve the speaker's definition as a performative. Eckhardt believes that the speaker as the creator of an utterance has the power to define the category of their creation. She implies that this is a general principle for acts of creation and suggests that the pragmatic principles at work are analogous to a painter's definition of their paintings' meaning. The example she gives is a depiction of a frog. Frogs have no visible features distinguishing between both sexes and therefore a picture of a frog would not specify its sexual category. According to Eckhard, the painter can define their painting as showing, say, a female frog, which would then be a specification of the frog's category and so change the interpretation of the picture. The painting is still interpretable without this definition but would not necessarily show a female frog. Unlike in the art case, where the definition is an explicit specification, for performatives Eckhardt proposes that without evidence to the contrary, the hearer assumes that the speaker is making the definition.

The analogy to graphic semiotics is based on the assumption that a painter creates a sign with a certain meaning but that their definition is conveyed via another medium, which in this case is language. If there is such an act of meta-communication for performatives, it is not (overt) in language. Eckhardt assumes that it is carried out implicitly, meaning that whenever an utterance is interpretable as performative, without evidence to the contrary, the speaker's definition is implied. This seems to be in line with Searle's assumption that the manifestation of the speaker's intention to perform a speech act is needed in order to do so. However, Searle assumes that the speaker's intention is manifested in the lexical semantics of performative verbs in combination with the self-referentiality of performative sentences. Considering self-referentiality a property of utterances instead of sentences is the only restriction under which I agree to this point of view, while rejecting an additional definition by the speaker.

In language, the category of an expression is usually related to its form or content. In order for a theory of performativity to fit into greater theories of language, Eckhardt's assumption of an extra-compositional definition by the speaker would constitute a rare and unsystematic exception. This would neither be theoretically elegant nor reasonable. Moreover, it is not necessary: Searle's account of intention manifestation can explain why (7) is not performative although it is self-referential: it has no performative verb which could express the intention to perform a speech act. Eckhard supports her supposition of speaker's definition with the example (9), in which a non-sincere utterance of a performative sentence is interpreted as non-sincere.

(9) *B: (gasps) Stop it! You are killing me!*
A: (laughing): Ok. I hereby promise to never be funny again.

Eckhardt argues that the context of the utterance, the mimic and gesture of the speaker in (9) constitute evidence enough for the hearer to assume the absence of *A*'s definition of their utterance as performative. The interpretion of *A*'s response as 'insincere'is however not due to the violation of linguistic requirements for performativity, but is based on the interpretation as a joke. Forms of figurative speech like irony, sarcasm or jokes flout a conversational maxim in Gricean terms and operate on the illocutionary force of an utterance. An assertive response to *B*'s utterance would not be taken seriously in a similar way:

(10) *B: (gasps) Stop it! You are killing me!*
A: (laughing): Ok. I know I am a bad person.

The arguments and the analogy supporting a *define*-step seem invalid, which leads me to disregard Eckhardt's pragmatic story. However, I am adopting her context-dependent account of self-referential utterances, and extend it in order to explain, what kinds of subjects and objects can be used in order to enable a performative interpretation.

3 The performative formula

3.1 Extending Eckhardt

Eckhardt's formal account of self-referential utterances has a lot to offer for a theory of performativity and speech acts, but given the non-sufficience of self-referentiality, how is performative self-verification derived? I suggest that an answer can be approached by investigating some implications of Eckhardt's logic of self-referential utterances:

We get from Searle that performative verbs lexicalise speech acts and performatives are a special case of speech acts in that they are realised explicitly. This means that the event referred to by a performative verb is a speech act. Eckhardt's analysis of self-referential performatives involves reference of the performative main verb to the communicative event, which implies the identity of a speech act and its utterance. This is crucial, because if they are the same event, they should have unique roles, arguments and aspectual and spatio-temporal properties. Austin's classic distinction between locutionary content and illocutionary force may hold for implicit speech acts, but collapses under this interpretation of performativity. Here, the locutionary content (i.e. the expressed proposition) describes the illocutionary force, while it conveys it. The locutionary and the illocutionary event not only coincide in performative utterances, they are the same event. Taking further this event-based account of self-referential utterances may provide us with a better understanding of performativity. A way to do this is to analyse the event structure of performatives in terms of arguments, thematic roles or aspectual type.

3.2 Performative event participants

Eckhard mentions the uniqueness of roles as a principle of event individuation in order to make her point that the self-referentiality of utterances can be established as being part of an abstract, more complex communicative event. If the speaker is the agent of the communication, they should be the agent of the performative event. The same point can be made for the hearer as communicative undergoer or recipient. The identity of the performative event with the communicative event entails that they have the same properties and roles, that the performative event is anchored in the same context as the communication (or utterance). It seems to be a viable assumption that speech acts have this restriction in general: that the participants of the locutionary event (communicative event) have to be the participants of the illocutionary event (speech act) as well and pass on their thematic roles. Concerning explicit speech acts, this should have an observable effect on the arguments of performative verbs, which can be formulated as a restriction on their reference. (11) illustrates that arguments of performative verbs require to be anchored in context in order for a self-referential interpretation of the utterance to be possible.

(11) a. *I (hereby) thank you.*
  b. *Thank you.*
  c. *The author of this paper hereby thanks her readers.*
  d. *Lisa (hereby) thanks Daniel.*
  e. *My employer (hereby) thanks you for your patience.*

If I uttered one of the sentences in (11a – 11c) to you, that would constitute an act of thanking. In Eckhardt's terms, this involves a simple communicative event between two parties (me and you) and therefore produces directly self-referential utterances. Not all of these sentences involve a first person subject, but something closely related: coreference of the performative agent with the speaker.

A performative interpretation of (11d) is possible only if uttered by Lisa to Daniel. There are no such contextual restrictions for (11c). Its verbal arguments are realised as definite descriptions, which again involve some deixis and therefore context-anchoring. In a written context, they have the same extension as the verbal arguments in (11a). This illustrates how third-person arguments can be part of the construal of performative meaning under certain contextual conditions.[2] The deictic verbal arguments in (11a + 11c) make explicit their contextual anchoring, whereas the descriptive arguments in (11d) presuppose coreference with speech participants under a performative interpretation. The sentence (11e) would need some additional context: at least it needs an authorisation for me to speak on behalf of my employer. Only in virtue of this circumstance, (11e) can be uttered performatively. This is no exception to the requirement that the performative agent has to be the agent of

---

[2] An anonymous rewiever pointed out that (11c + 11d) are not as straightforwardly acceptable as performatives, with deictic arguments. This is also noted by Eckhardt (2012), who assumes that third-person-subject performatives require explicit context-anchoring through *hereby*. They however occurr, especially in written language. Third-person realisations of performative participants have different functions in relation to their context-anchoring: I assume that deictic third-person arguments as in (11c) are chosen as manifestation of a rather formal register. A weak definite variant (*the author(s)*) is however more commonly used than constructions with possessor specification. Non-deictic third person subjects may be used in order to specify the identity of the communicative/performative agent, which might not be salient in written communication at all times. A non-deictic third person object like in (11d) can be a means of domain restriction, which operates on the set of communicative recipents and singles out the intended performative repicient(s). This is especially common in spoken or written communication which is distributed to multiple recipients. Just think of people saying things like *'I hereby greet my mother.'* on television.

the communicative event referred to by the performative verb. The event of thanking in (11e) is a larger communicative event between my employer and you, which is relayed via my utterance. Because my utterance is an integral part of this larger communication, it establishes an indirect, mereological self-reference.

Under a self-referential interpretation, the participants of the thanking-event in (11) have to be anchored in context: in order for the performative verb *thank* to refer to the communicative event, its agent is linked to the speaker and its recipient is linked to the communicative undergoer. This is predicted by Parson's argument that event participants and their roles are constitutive for events and Eckhardt's event-based analysis of self-referentiality: if the performative event is the utterance, also its constituent parts have to be parts of the utterance situation. The context-anchoring requirement of performative events and their subparts contributes to a description of how communicative events and interlocutors are conceptualised and realised in language. This is derived from and parallel to Eckhard's self-reference restriction for event arguments of performative verbs and the different ways in which it comes about.

The old performative formula explained

The use of proximal deictic expressions is an explicit realisation of context-anchoring of verbal arguments. It is therefore not surprising that a first person agent, second person undergoer and present tense are so common among performatives. The semantic composition of performatives with explicitly context-anchored arguments is modeled parallel to Eckhardt's analysis of performatives with the deictic adverb *hereby*:

(12) a. $[\![\textit{I}]\!]^{w,c} = \mathbf{sp}$ (speaker in c)

b. $[\![\textit{you}]\!]^{w,c} = \mathbf{h}$ (hearer in c)

c. $[\![\textit{hereby}]\!]^{w,c} =$ ε (communicative event in c)

d. $[\![\textit{thank}]\!]^{w,c} = \lambda y \lambda e \lambda x.\textsc{Thank}(x, e, y)$

e. $[\![\textit{I hereby thank you.}]\!]^{w,c} = \textsc{Thank}(\mathbf{sp}, \varepsilon, \mathbf{h})$

f. $[\![\textit{I thank you.}]\!]^{w,c} = \exists e.\textsc{Thank}(\mathbf{sp}, e, \mathbf{h})$

The argument slots of the THANK-predicate in (12e) are saturated with deictic expressions, explicitly realising their context-anchoring. Note that context-anchoring of the participant-arguments is compatible with explicit self-reference of the event-argument and the arguments have to be compatible for a successful interpretation. This is for a relationship between the event and its participants, which suggests that the event and its participants have no equal status as arguments. The reference of the event argument and the reference of the participant arguments depend on each other.

Existential binding and cirumstantial coreference

Realising arguments as specific existential statements is not exclusive for event-arguments. Eckhardt brings up a specific existential binding of the subject of *'Someone needs a bath here.'*, which you could perfectly imagine if uttered by a mother to her son. Also, conventionalised omissions like in (13) are not unusual:

(13) $[\![\textit{Thank you.}]\!]^{w,c} = \exists x \exists e.\textsc{Thank}(x, e, \mathbf{h})$

A performative utterance of (13) necessarily involves self-referentiality and therefore context-anchoring of all verbal arguments. That the communicative undergoer is explicitly realised as the performative undergoer is compatible with that interpretation. Of course, a large dose of social convention plays a role for determining the preferred interpretation of such existential statements. (13) is one of the most frequently used performatives, which is probably a factor, which made the conventionalisation of this omission possible in the first place and thus ensured that contextually anchored reference is the associated interpretation.

The way in which participant-arguments depend on the event-argument explains why (intended) performative sentences with third person subjects are often infelicitous. Third person NPs can refer to persons other than speech participants. But also, although it is unusual, they can refer to interlocutors, which presents one of the advantages of an event based account of performatives: The felicitousness of third-person agent performatives is no longer puzzling.

(14) a. $[\![\textit{John hereby thanks Mary.}]\!]^{w,c} = \textsc{Thank}(\mathrm{J}, e, \mathrm{M})$

b. $[\![\textsc{Thank}(\mathrm{J}, \varepsilon, \mathrm{M})]\!]^{M,\, g(\mathrm{J}/\mathbf{sp},\, \mathrm{M}/\mathbf{h})} = 1$

Although there are no existential statements in(14), performative context anchoring is established similarly here: If the subject refers to the speaker and the direct object refers to the hearer, a performative interpretation is possible. Because the reference of the event-argument is explicitly specified through *hereby*, this is the only possible felicitous interpretation. Third person arguments do not hinder a performative interpretation in general, but only when they refer to a person who is not a speech participant (which is mostly the case). In English, the third person is not deictic, like the first and second person are. While the first and second person specify the role of an NP in the communicative event, the third person has no such specification.[3]

Mereologically self-referential utterances

Certain social conventions (e.g employment, legal representation) allow persons to communicate on behalf of others. A sender A communicating with a recipient B via a messenger C gives rise to a complex communicative event with smaller communicative events as proper subparts. Eckhardt motivated her generalisation of performative self-reference as reference to an abstract communicative event with (4a), an example of a sentence, which could be uttered as the temporally ultimate subpart of the complex communicative event: the communication between C and B.

(4a) *King Karl hereby promises you a cow.*

A felicitous utterance of (4a) constitutes an exception to the principle that performative arguments refer to immediate interlocutors. As Eckhardt points out, this and similar cases involve an indirect kind of eventive self-reference, which is why they allow for an indirect context anchoring of participant-arguments. The indirect self-referentiality of a performative utterance u by C towards B on behalf of A comes about through reference of the event argument to the larger communicative event c between A and C. This is no strict self-reference of u, but because $u \subset c$, it is an indirect kind of self-reference, which can be described as mereological. The relationship between the event-argument and the participant-arguments stays the same: the participant-arguments of the verb have to be the participants of the event. Therefore, the performative agent as expressed in the utterance has to be the communicative agent A. A felicitous utterance of (4a) also presupposes some *'authorised-to-speak-on-behalf-of*-relation between the speaker C and the performative agent A. Only in virtue of this relation, it can be a felicitous performative. The context-anchoring of performative arguments is met as a (less strict) relational association of performative participant-arguments with interlocutors. This relational association can also be made explicit through the use of relational nouns, possesive constructions or weak definites[4]. The possessive constructions with first-person possessors in (15) are therefore an explicit realisation of associative context anchoring

---

[3] This is different in languages with obviative marking, like for example some Algonquian languages. They specify the role of a third person with respect to the utterance context as proximate or obviative. This analysis predicts that obviative realisations of participant-arguments should not be allowed coreference with speech participants. If that is the case, they should not allow for an interpretation as strictly self-referential performatives. They might, however be allowed in performatives which are conveyed on behalf of others, as they are less restricted.

[4] Cf. Löbner (2011) for an account of nominal relationality and different ways in which it comes about. It is based on theories which assume an associative structure and a subcategorial concept type as part of the lexical semantics. The semantic features ± uniqueness and ± relationality are assumed inherent to lexical nouns and their cross-classification gives rise to a four-way distiction of nominal concept types.

(15) a. *My employer (hereby) thanks you for your patience.*
b. *I request payment from your client.*

Although a variant of (15a) with a non-relational subject (e.g. *The Café du Congo*) can be uttered performatively (e.g if the Café du Congo is my employer), a relational subject is probably a frequent choice for such relayed performatives, because it explicitly expresses the relation between the messenger and the communicative agent.

4 Conclusion

What does it mean for an utterance to be self-referential? In Eckhardt's terms it means that the event argument of the main verb refers to the ongoing act of information transfer. Self-reference can be explicitly realised or implicitly achieved (with and without *hereby*). It can also be direct (when the information transfer is established on utterance-level) or mereological (involving a superordinated complex information transfer). In a Davidsonian account, the event denoted by a verb is formalised as argument, while other verbal arguments realise the participants in the denoted event. If the event is anchored in context, its participants will have to be context-anchored as well. This, in turn, can also be explicitly realised or implicitly achieved (with and without deictic expressions/relationality). The interdependence between event and participants comes about, because the participant-arguments have a role in the event. This raises the question if the different kinds of verbal arguments have a different status, which should be subject to further research. For now, it explains why verbal arguments which are not anchored in context lead to:

1. Unavailability of a self-referential interpretation for *hereby*-less sentences.
2. Infelicitousness of sentences with *hereby* due to incompatible participant-event-combinations.

The self-referentiality of an utterance is necessary for it to be performative, but not equivalent to self-verification, another necessary condition. So how is performative self-verification derived? Searle assumes that it comes about about through composition with a performative main verb in combination with self-referentiality. Performative verbs can be used in descriptive sentences and their potential to manifest an intention to perform a speech act is only realised in combination with self-reference. I showed that self-referentiality can not only be understood as contextual anchoring of the event-argument, but also entails context-anchoring of its subparts. This, combined with verbal meaning can be understood as a link between self-reference and self-verification. The non-self-referential *'I am here now.'* is self-verifying because of its composition, the meaning of its main verb and the contextual anchoring of its subparts. Performative self-verification seems to be achieved in the same way.

An event-based account of self-referential utterances is substantially connected to the semantics and pragmatics of dialogue:

Performative meaning can only be interpreted in the context of the communication or dialogue it occurs in. This is a consequence of accounting for context-anchoring of performative event-participants as coreference with speech participants or relation to speech participants, respectively. It is also a consequence of the identity of the performative event with the communicative event. The other side of the coin is that communicative events can be directly referred to by performative verbs, therefore studying performatives enables us to directly observe how language treats them. One thing, that Eckhard's account tells us, is that communication is not always carried by a single utterance event, but can be conveyed via people communicating on behalf of others. In that case, several communicative events with different participants contingently form an overarching communicative event.

References

John Langshaw Austin. Performative utterances. In *Philosophical Papers*, pages 233–292. Oxford University Press, 1979.

Kent Bach. Performatives are statements too. *Philosophical Studies (Minneapolis)*, 28(4): 229–236, 1975.

Donald Davidson. The logical form of action sentences. In Donald Davidson, editor, *Essays on actions and events.*, pages 105–122. Clarendon Press, Oxford, 1980.

Regine Eckhardt. 'hereby' explained: an event-based account of performative utterances. *Linguistics and Philosophy*, 35:21–55, 2012.

Sebastian Löbner. Types of nouns, NPs, and determination. *Journal of Semantics*, 28:279–333, 2011.

Terence Parsons. *Events in the semantics of English. A study in subatomic semantics.* MIT Press, Cambridge, 1990.

John Searle. How performatives work. *Linguistics and Philosophy*, 12(3):535–558, 1989.

Timing and Grounding in Motor Skill Coaching Interaction: Consequences for the Information State

Julian Hough^{1,2}, **Iwan de Kok**^{1,2,3}, **David Schlangen**^{1,2}, and **Stefan Kopp**^{2,3}
^{1}Dialogue Systems Group, ^{2}CITEC, ^{3}Social Cognitive Systems Group
Bielefeld University
julian.hough@uni-bielefeld.de

Abstract

While tutorial dialogues have been well-studied, the nature of dialogue in physical coaching scenarios is much less well understood. We present a corpus study on coaching interactions wherein a coach trains a trainee to improve a motor skill. We show how our findings put novel requirements on pedagogic dialogue act taxonomies, grounding criteria and information state update models of situated dialogue. One of these requirements is to distinguish between grounding in the traditional sense along an understanding dimension and grounding in terms of a motor program schema, the latter being due to the coach's goal to transfer knowledge of a physical movement to the trainee. Another requirement is that a fine-grained notion of time, both in absolute and relative terms, must become a first class citizen of the dialogue state to be able to model motor skill coaching. A final requirement for an information state model is characterizing what is under discussion and in the established common ground– in these kind of domains this is generally not questions and propositions, but skills and their desired and observed parameters.

1 Introduction

Dialogue in pedagogic domains presents interesting challenges for corpus studies and formal dialogue models. In contrast to more commonly studied task-completion oriented dialogues where an instructor influences their instruction follower's action towards a successful outcome (i.e. to do something), in pedagogic domains the intended outcome for the instructee is *learning gain*– that is, measurable improvement at the task at hand (i.e. to learn *how* to do something or improve upon it).

Dialogue research in tutorial domains requires a relevant dialogue act (DA) taxonomy that deals with grounding understanding of a given skill or piece of knowledge, such as those in (Boyer et al., 2007; Boyer et al., 2008; Boyer et al., 2009). The most well-developed DA taxonomies designed for task-completion dialogues (e.g. DAMSL (Core and Allen, 1997)), while sharing certain communication management DAs, require extensions with DAs that capture 'know-how'– that is, the transmission of skill, knowledge and technique from tutor to tutee. The nature of the *feedback* on tutee attempts, both successful and unsuccessful at the task at hand is also crucial to the taxonomy. Particularly, the degree of positive affect with which a tutor gives feedback can influence learning outcomes. Tutorial dialogue systems such as (Litman and Silliman, 2004; Graesser et al., 2005) use the insights from DA-based corpus studies in their systems to generate appropriate dialogue acts to maximise learning gain.

Situated dialogue, where participants are either physically co-present or have access to a commonly shared virtual space, presents other challenges for DA taxonomies. Grounding DAs such as feedback and repair need not only reference previous verbal utterances, as in (Schegloff et al., 1977), but can also reference non-verbal actions which concern a physical task at hand. In this regard Raux and Nakano (2010) study three types of non-verbal action corrections in a computer-game dialogue whereby a manager guides a player through a task in a virtual environment. Failures in communication are addressed by the manager via correction of errors of three observed types: *Commission* (failure to do the expected or appropriate action), *Omission* (failure to react to an instruction) or *Degree* (appropriate type of action carried out but falling short of the intended outcome by some real value). They showed the three correction types were uniformly distributed, but showed differences in timing: Commission and Degree corrections were likely to be produced much closer to the error-containing action's start time (on average 2.3 seconds and 2.4s) than normal non-corrected instructions were to correct actions (3.8s).

Embodied situated dialogue becomes even more complex to analyse when gesture and speech interact. Lücking et al. (2013) provide a rich gesture taxonomy and morphology with a proposed interface to the semantics of speech. While this mark-up is comprehensive, it focuses on the transfer of spatial scene descriptions. The domain is face-to-face route description whereby a route giver will make frequent use of iconic and deictic gestures to indicate where their dialogue partner should be locationally during their route. This is a complex type of multi-modal knowledge, but not the embodied procedural knowledge required in motor skill learning which we focus on here.

In this paper, we address the intersection of pedagogic and embodied situated dialogue found in the domain of motor skill coaching, and describe findings which reveal part of the nature of the information state of the participants in these interactions. The rest of the paper is as follows: §2 describes the challenge and uniqueness of motor skill coaching dialogues, §3 outlines our research questions, §4 describes our findings on timing and grounding in a coaching corpus study, §5 describes the consequences for information state update approaches to dialogue modeling and §6 concludes.

2 Motor Skill Transmission in Coaching

For a technical skill such as computer programming, learning gain can be assessed by tutors and tutorial systems by generating open questions about procedural knowledge such as "What should you do now?" to which the tutee can provide an answer (e.g. "I will use an array") to show evidence of their competence (Boyer et al., 2008). The tutor evaluates the tutee's progress by such question answering, and gives appropriate feedback. The goal is fairly clearly presented in this cases to the tutee as the learning gain criterion is set out in advance– for instance the achievement of higher scores.

However, in motor skill learning, for a human coach, task success is much more difficult to evaluate and communicate, particularly if the outcome is not directly observable by the trainee. In such purely technique-oriented tasks, the feedback from a coach is vital to learning success, and for novices, this feedback defines it.

Furthermore, under McMorris (2014)'s definition of a physical skill as "the consistent production of goal-oriented movements, which are learned and specific to the task", the requirements beyond factual knowledge learning increase again: the situated, embodied nature of a motor skill means feedback both from the coachee's own perceptual self-monitoring and externally from the coach is time-critical, with online instructions being of utmost importance.

We will assume the coach's goal is to induce in the trainee a *motor program schema* (Schmidt, 1975), and evidence as to whether the coachee has induced it or not is observed through their demonstration of the desired outcomes. The feedback on successful learning is relayed to the coachee to ground the fact it was successful.

Two types of grounding: understanding and skill To model motor skill coaching interactions, we propose there are two types of grounding at work– grounded$_{understanding}$ and grounded$_{skill}$. For a skill to become grounded$_{understanding}$, it has to be subject to communicative grounding requirements in the spirit of (Clark, 1996). However, this domain centers around communicating non-propositional information of physical movement which is only observable by *consistent demonstration of success* by the coachee– only then, after positive feedback by the coach will this become grounded$_{skill}$.

The reason we make this division is that it is possible the coachee could resolve all linguistic and intentional information in a description of an exercise but still not have grasped the skill, either in kind and in degree. The physical, embodied nature of learning a motor program schema means this representation is not straightforwardly translatable into symbolic means for information transmission but needs analogue values for trajectories, speeds, distances and pressures. It is clearly challenging for a coach trying to make this information common ground, both in terms of the dialogue acts and nonverbal actions they use and the *timing* they employ to do this.

3 Research questions

To investigate timing behaviour and grounding strategies in coaching interactions we conduct a corpus study which focuses on the following overarching research questions:

q1 **Characterizing dialogue acts for motor skill coaching:** What is an adequate taxonomy of dialogue acts and non-verbal actions for the motor skill coaching domain, and how does it differ to existing task-oriented, tutorial and situated taxonomies?

q2 **Modelling dialogue context:** What type of information is in the *dialogue context* for the coach during the current coaching interaction as it unfolds in terms of the skill elements addressed so far? When does the coach take skills introduced to be understood by the trainee ($\text{grounded}_{understanding}$) and when does the coach take a skill to be part of the coachee's motor program schema (grounded_{skill})?

q3 **Modelling decisions:** Which elements of the context influence the type of dialogue act the coach will use to address it? Specifically, does the status of the skill element as *given* ($\text{grounded}_{understanding}$) or *new* affect the dialogue act type, and does the status of the skill element in the common ground as having been routinized and mastered (i.e. being grounded_{skill}), affect the way the coach talks and acts non-verbally concerning the skill element?

q4 **Timing:** When do dialogue act and non-verbal actions happen with respect to coachee's skill attempts on a fine-grained time-line? As timing is a critical part of motor skill acquisition this becomes more vital than other tutorial domains.

4 Corpus Study: Timing and Grounding in Coaching Dialogue

To address the research questions we study a corpus of coaching dialogues where a coach trains a trainee in the exercise of a body-weight squat (a squat done with no weight or barbell)– see Figure 1. It is a simple and closed skill in that it is not subject to environment change (i.e. it is not an interactive sport) and can be practiced alone. However, it is an interesting skill from a dialogue perspective in that it is an exercise without a tangible outcome (such as scoring a goal in practicing taking football penalties), and relies on the expertise of a coach to provide feedback to indicate success.

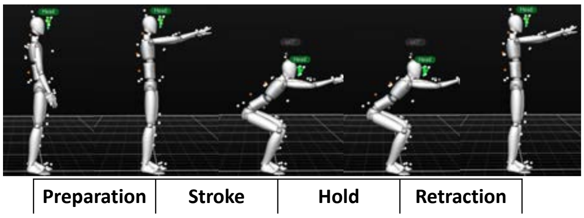


Figure 1: The four phases of a squat

We invited 8 participants to interact with 2 different professional fitness coaches (4 participants per coach). The average length of the sessions was approx. $4\frac{1}{2}$ minutes. The participants had various different levels of expertise with squats ranging from novice to doing it on a monthly basis. None were professional athletes but all partook in recreational exercise.[1] All sessions were in German and all participants were native German speakers.

4.1 Dialogue act and non-verbal action annotation

The dialogues were transcribed, translated and utterances were segmented into dialogue act units.[2] To address question q1 we did an initial analysis of 2 sessions, one from each coach, and created our annotation scheme for verbal and non-verbal acts in Table 1. The verbal dialogue acts specialized to this domain are as follows:

- *Instruction*[*directive*]: Imperative command to carry out a skill[3] (e.g. "Do three or four squats")
- *Instruction*[*attempt*]: Request to carry out a skill to the best of the participant's ability (see e.g. (1)).

(1) **Coach:** also langsam so weit runterarbeiten .. wie du runterkommst
so slowly go down .. as far as you can

- *Instruction*[*mentalize*]: Imperative to imagine something not present that will help with the skill, or to pay attention to the feeling of a particular part of the body during skill attempts (e.g. "Imagine there is a wall in front of you and you do not want to touch it")
- *Acknowledge*[*skill*]: Signal of recognition of a skill attempt with neutral sentiment, analogous to standard backchannels (e.g. 'Right' or 'Okay' said after a squat has been completed)

---

[1] We do not look at the effect of expertise or prior experience here, but intend to in future work.

[2] Like the slash-unit of (Meteer et al., 1995).

[3] This is analogous to the Action-Directive in (Core and Allen, 1997).

- *Adjust*: An instruction where the *degree* of an element of the skill is directed to be changed, as in (2). This is similar to (Raux and Nakano, 2010)'s Degree Error Correction, however an *Adjust* has a different notion of success– in a task-completion dialogue such as object selection there are binary notions of success and failure and reinforcement of good practice is not vital, whereas here the reaction from the corrector is especially important in terms of its motivational affect and long-term learning outcome.

(2) **Coach:** Stell mal die Beine etwa schulterweit auseinander ...
also noch [ei]nen Hauch weiter
Plant your feet about shoulder-width apart .. a bit more

- *Repair*[*skill*]: An other-repair of a skill attempt which repairs misunderstanding of the intended outcome (rather than a linguistic other-repair repairing an utterance) or recovers from a lack of uptake via repeating or reformulating (e.g. "No, not that way, the other way").[4]
- *Explanation*: An explanation of why a certain skill is important (rationale) or consequences of mastery (e.g. "this will help your power transmission"), or an elaboration on an instruction with more descriptive detail (clarification).[5]
- *Feedback*[*positive*]: Evidence of approval that the skill is being performed well (e.g. "You kept your back nice and steady").
- *Feedback*[*negative*]: Criticism of the way the skill is being performed (e.g. "At the moment your knees are buckling a lot").
- *Commentary*[*self*]: Commentary description on the current action by the speaker (e.g. "I'm now tensing my stomach and back muscles").
- *Commentary*[*other*]: Commentary description on the current action by the addressee (e.g. "You're now getting into what we call the neutral position").
- *SetGoal*: Announcement that the session will turn its attention to a given skill element (e.g. "Let's focus on the width of your stance").

The other acts shown in Table 1 have standard definitions for dialogue act tags. The non-verbal acts specialized to this domain are:

- *SkillAttempt*[*preparation*|*stroke*|*hold*|*retract*]: An instance of an attempt at a phase of the overall target skill. For squats the 4 phases are as in Figure 1. The *preparation* phase can consist in several parts from adopting the stance to raising or crossing the arms. The *stroke* is the main phase focused on by our coaches. The *hold* at the lowest position is often short and occasionally too short to annotate at all. The quality of the squat has largely been determined before *retraction* back to the upright position.
- *Demonstration*[*exaggerated*|*positive*|*negative*]: Presentation of a movement either as it is meant to be done (*positive*), else an example of what not to do (*negative*). This can be *exaggerated* to emphasize an element of the skill.
- *Iconic*[*modelling*|*shaping*]: Gestures which represent objects, which are invariably parts of the body involved in the skill, either through using other body parts to represent them (*modelling*) or *shaping* their outline in the air.
- *Deictic*[*self*|*other*|*thirdperson*|*touch*]: Gestures used to refer to something in the environment. These include touching of the body in this domain, both one's own (*self*) or one's partner's (*other*) to point out physical details of movements– see Figure 2 C for an example of a $Deictic[self]$ gesture.

The other non-verbal acts are *Beat* gestures, *Head* movements (including nods) and *Discourse* gestures. The category *OtherAction* includes concurrent movement of the participants around the experimental space.

Skills Under Discussion In addition to these acts, for each act decision the annotators chose the particular element(s) of the motor skill being talked about. We will call these tags the *Skills Under Discussion*– their relationship to Questions Under Discussion models discussed in Section 5. The labels form a closed set and consist of values such as $StanceWidth$, $ArchedBack$ and other squat-specific skills. These approximate the *content* of the acts in this domain.

Annotation agreement and overall distributions Three annotators annotated the corpus and

---

[4] This is similar to the Commission and Omission errors in (Raux and Nakano, 2010) described above.

[5] We encourage the attributes [rationale] and [clarification] to be added to the tag where annotators are confident which one it is, however we do not calculate agreement levels on this.

we checked annotation agreement on verbal dialogue acts only for one transcript between two of them. We had an acceptable Cohen's κ of 0.69. The main source of disagreement was over what constituted *Commentary[other]* and what was *Feedback[positive]*– the boundary between these can be vague upon manual inspection.

As can be seen from the distributions in Table 1, as expected the verbal side of the interaction is heavily dominated by the coach. The majority of the coachee's verbal contributions were acknowledgements in the form of short backchannels indicating understanding of the coach's dialogue act.

30.0% of the coach's DAs are directive instructions, and adjustment acts such as in Figure 2 D are relatively common (81 total, 9.1% of the coach's DAs). Positive feedback is overwhelmingly more common than negative feedback (112 occurrences vs. 18) and the acknowledgement signal $Acknowledge[skill]$ that a skill has been seen by the coach is also frequent (11.1% of coach DAs). We will discuss grounding strategies below.

For non-verbal acts, there is again asymmetry in the distributions, which is unsurprising given the domain– the coachee attempts a skill whilst the coach demonstrates it. What is more interesting however is the frequent use of iconic and deictic gestures, together constituting over 40% of the coach's non-verbal actions.

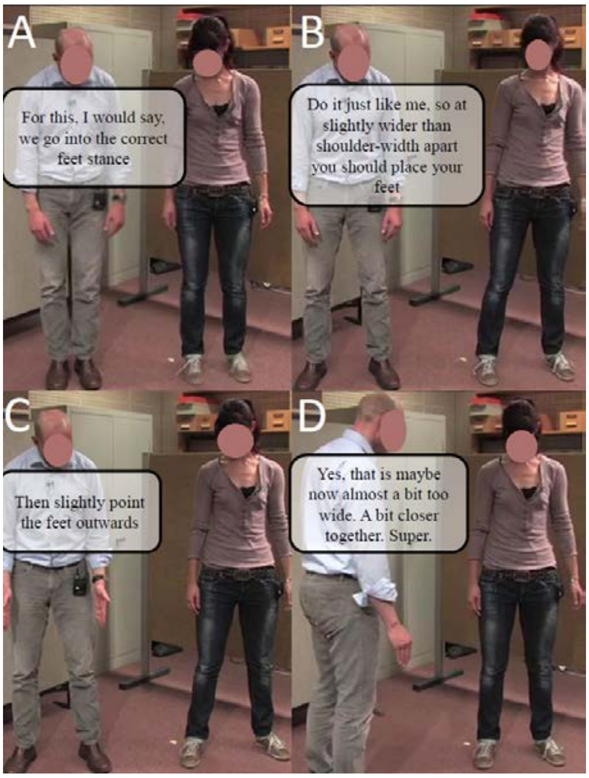

A
For this, I would say, we go into the correct feet stance
B
Do it just like me, so at slightly wider than shoulder-width apart you should place your feet
C
Then slightly point the feet outwards
D
Yes, that is maybe now almost a bit too wide. A bit closer together. Super.

Figure 2: A typical coaching interaction: In A the coach commentates on where the foot position should be. In B the coach elaborates on the instruction. In C he uses a deictic gesture relative to his own body to show the correct width and in D he repairs the coachee's over distance and *adjusts* her stance until satisfied.

4.2 Timing of Adjust and Instruction acts

Having gained an insight into the interactions through dialogue acts, we now focus on the timing of these acts, and in particular, the timing of a coach's dialogue act production relative to a skill attempt by a coachee. As just described, *Adjust* moves are very common and have interesting timing properties in terms of turn-taking– see Figure 3 for a time-line of an adjusting event. The coach, constantly monitoring the coachee's action keeps incrementing his contribution with the adjunctive phrase 'noch ein bisschen' ('a bit further') until the coachee has achieved the desired foot stance. Notice the timing here is incredibly fine-grained, with the coachee's reaction being close to human reaction time (≈0.2s) from the *middle* of the adjust instruction. Adjustments are inherently able to be concurrent with the non-verbal channel of the coachee's action, so tight coupling of the coachee's movement and the coach's feedback, although appearing like a normal dialogue turn-taking structure on first pass, can afford a great deal more overlap.

In non-adjustment instructions, timely reaction from both parties is also common. In fact, we observe coachees often *anticipate* the instructions even in these forward looking acts. To investigate this observation empirically we calculated the mean and standard deviations for the time between the end of an $Instruct[directive]$ act and a skill attempt corresponding to the phase of the squat it is instructing. We do this both for instructions to enter the stroke phase such as 'geh nochmal runter' ('go down again') and for the retraction instructions like 'komm wieder hoch' ('come up') and find the means (vertical red lines) and probability density plots as shown in Figure 4. We find the mean interval from the end of the instruction to the start of the skill attempt was negative for the stroke phase at -0.274s (st.d.=1.204) and even more so for the retract phase at -0.410s (st.d.=0.510), meaning on average the coachees

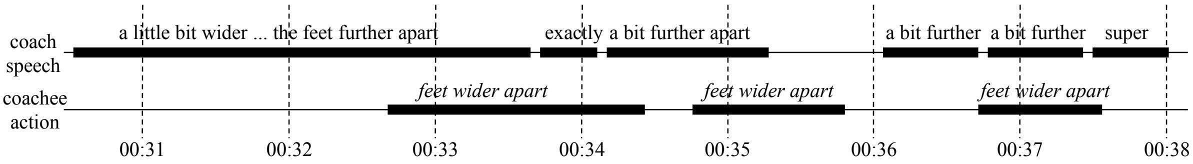


Figure 3: The fine-grained interaction between the coachee's actions and the coach's adjust moves

	Count (%)	
Dialogue Acts	**Coach**	**Coachee**
Instruct[directive]	267 (30.0)	1 (0.5)
Instruct[attempt]	41 (4.6)	0 (0.0)
Instruct[mentalize]	20 (2.2)	0 (0.0)
Acknowledge[skill]	99 (11.1)	0 (0.0)
Adjust	81 (9.1)	0 (0.0)
Repair[skill]	10 (1.1)	0 (0.0)
Explanation	80 (9.0)	0 (0.0)
Feedback[positive]	112 (12.6)	3 (1.4)
Feedback[negative]	18 (2.0)	0 (0.0)
Commentary[self]	32 (3.6)	8 (3.8)
Commentary[other]	2 (0.2)	0 (0.0)
SetGoal	17 (1.9)	0 (0.0)
Acknowledge[verbal]	32 (3.6)	150 (70.4)
Question	31 (3.5)	4 (1.9)
Answer	3 (0.3)	22 (10.3)
FloorManagement	21 (2.4)	2 (0.9)
StatementOther	11 (1.2)	17 (8.0)
Social	9 (1.0)	2 (0.9)
ClarificationRequest	5 (0.6)	4 (1.9)
Non-verbal Acts		
SkillAttempt	1 (0.2)	398 (75.2)
Demonstration	132 (21.5)	0 (0.0)
Iconic	134 (21.9)	1 (0.2)
Deictic	114 (18.6)	2 (0.4)
Beat	39 (6.4)	0 (0.0)
Head	25 (4.1)	28 (5.3)
Discourse	45 (7.3)	21 (4.0)
OtherAction	123 (20.1)	79 (14.9)

Table 1: Dialogue acts and non-verbal communicative actions marked up in our corpus with the percentage of total acts for each tag.

were moving on the Skill Under Discussion well before the end of the utterance. The coachees can take initiative and predict instruction completions easily, just as initiative is possible in other domains as described in Traum and Larsson (2003).

4.3 Grounding the motor program schema

In the outset in §2 we suggested there are two principal grounding mechanisms at work in motor skill coaching. From the coach's perspective, they must ensure not only that the coachee understands the meaning of the current Skill Under Discussion (i.e make it grounded$_{understanding}$), but that they use it to induce the appropriate motor program schema knowledge, whereupon they should provide feedback that this has been done successfully (make it grounded$_{skill}$).

We observe the coaches use various techniques to achieve the second grounding criterion. While purely instructing the coachee through directives can be effective initially, they must use other techniques if difficulties persist. In all of our sessions, one or two Skills Under Discussion were addressed at much greater length compared to the others because they were problematic for that particular coachee.

We investigate the effect of grounded$_{understanding}$ status of a Skill Under Discussion, a status we assume by virtue of the fact it has been addressed before by the coach and that the coachee has performed it to the best of their ability, for which they received acknowledgement or even positive feedback. We find that when a skill is re-referenced verbally there is a difference in dialogue act type used. We calculate the distribution of dialogue act types used based on whether the skill has been openly raised before or is new– see Table 2. While $Instruction$ acts are the most probable in both first mentions and subsequent mentions, their dominance is attenuated in the subsequent condition. $Adjust$ moves are one such way to attune the parameters of a skill as discussed, but also $Explanation$ becomes more frequent, as does $Feedback[positive]$ and $Instruction[mentalize]$ instructions. The $Acknowledge[skill]$ act, while having similar lexical and phonetic qualities to normal backchannel acknowledgements (e.g. 'okay'), is a grounding mechanism where the coach

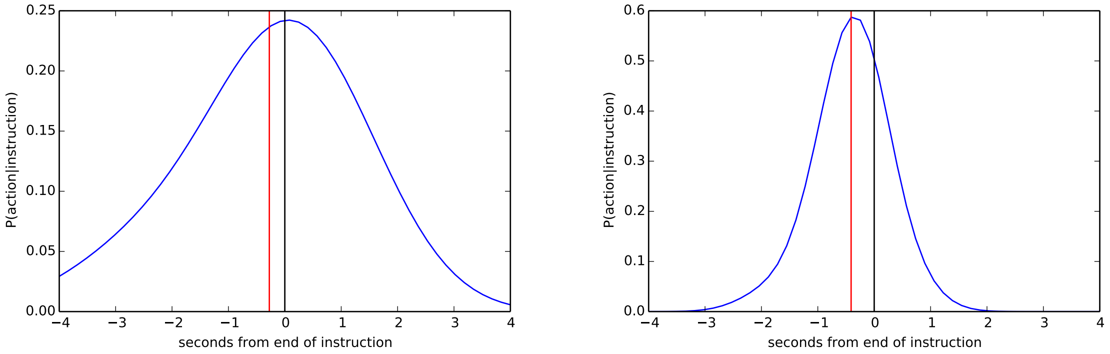


Figure 4: Anticipation in the uptake of an instruction both in the stroke (left) and retraction phase (right)

communicates a message to the effect ‘I’ve seen you attempt this’, however it is not strong enough evidence for the message ‘I’ve seen you master this’, and positive feedback is the way to convey this and make it grounded$_{skill}$.

There were a handful of mentalizing examples where the coach uses imagery and metaphor, such as in (3). These only occur in the longer sessions where a particular problem has been addressed numerous times.

(3) **Coach:** Versuch mal gedacht so ein bisschen Froschbeine zu machen das heisst wenn du runtergehst die Knie eher auseinanderzudrücke
Try to think about frog legs when your knees start getting closer together

In non-verbal behaviour, there are also differences with the gesture accompanying the dialogue acts which reference the skills. In explanations, not only are given skills likely to be accompanied by an overlapping gesture (87% new versus 97% given), also qualitatively there is a shift from *Deictic* gesture to *Iconic* gesture and *Demonstration*– see the bottom of Table 2. Analogously to the verbal case with direct instructions, directness through deixis is initially preferred to ensure grounding$_{understanding}$, but to achieve grounding$_{skill}$ several techniques of both personal demonstration and analogy with other objects and images is required.

5 Consequences for an Information State Model of Dialogue

We have argued it is useful to distinguish between grounding in the traditional sense along an understanding dimension and grounding a motor program schema, the latter being due to the coach’s goal to transfer knowledge of a physical movement to the trainee. We show skill elements behave similarly to discourse referents in that their given versus new status affects the dialogue act type with which they are re-referenced. This puts the requirement on an information state model of dialogue that what is under discussion is not al-

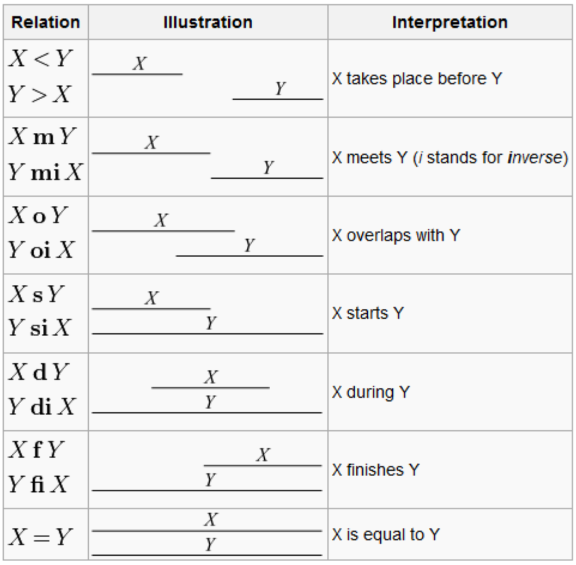
	% of acts about sub-skill	
Dialogue Acts	**1st ref.**	**Subsequent**
Instruct[directive]	46.2	25.9
Instruct[attempt]	17.3	4.6
Explanation	8.7	15.8
Feedback[positive]	6.7	11.8
Adjust	5.8	18.2
SetGoal	2.9	0.7
Question	2.9	2.4
Feedback[negative]	2.9	3.6
Commentary[self]	1.9	1.9
Instruct[mentalize]	1.9	4.1
Acknowledge[skill]	1.0	7.4
Commentary[other]	1.0	1.0
Repair[skill]	1.0	1.9
Non-verbal Acts		
Iconic	30.4	39.4
Deictic	26.1	15.2
Demonstration	8.7	15.2
Discourse	4.3	6.1
Beat	2.2	6.1
OtherAction	15.2	15.2
None	13.0	3.0

Table 2: Different dialogue acts and non-verbal acts used when a skill element is referred to the first time and subsequently. Note the non-verbal acts are only those overlapping Explanations here

ways propositional material, but internal representations of action sequences. Instead of issues being *resolved* like a QUD-based model (Traum and Larsson, 2003; Ginzburg, 2012), the coachee must *demonstrate* their acquisition of a motor program schema. While one could posit that demonstrations evidence propositions such as $CanDo(x)$ for a skill x, which 'answer' whether $CanDo(x)$?, bi-valued propositions may not be a useful analogy given the real values and degrees that need to be parameterized in skill representations.

Another requirement arising from our findings, and a general short-coming of the traditional Information State update approaches is the lack of timing information in the information state, which in real-time situated dialogue such as coaching dialogues is crucial. In situated multimodal dialogue interaction, the state needs to represent time to account for the plethora of overlap both inter-modally (speech and gesture occurring with various degrees of synchronization within the same agent's behaviour) and interactively (speech and gesture of different agents overlap or synchronise with one another to various degrees). In both cases the nature of the synchronization is important for meaning construction, a fact currently exploited more by the Virtual Agents community (Kopp et al., 2014) than by dialogue theorists and semanticists– however see Lücking et al. (2013).

One theoretical and practical step we are exploring is using an established temporal reasoning system, Allen's interval algebra (Allen, 1983), which describes the possible relations two temporal events can have to each other, with the primitives, or *base relations* as in Figure 5. According to the assumption of the classical information state approach, for two contiguous dialogue acts by two different agents which are related (i.e. a 'minimal pair' of dialogue acts) A and B, their relative timing would be represented $A < B$ or A m B (A ends completely before B begins, either with no gap or contiguously). However, we argue that if A was a 'forward-looking' move, such as an instruction, and B was a 'backward-looking' move related to A such as a skill attempt, all 13 Allen relations between A and B are possible, even $A > B$ and A mi B. To model coupling between two or more multimodal dialogue acts as shown here, an approach using the constraints of this temporal algebra permitting overlap and anticipation between acts and intra-act level increments is required.



Relation	Illustration	Interpretation
$X < Y$ $Y > X$	X Y	X takes place before Y
$X \mathbf{m} Y$ $Y \mathbf{mi} X$	X Y	X meets Y (*i* stands for *inverse*)
$X \mathbf{o} Y$ $Y \mathbf{oi} X$	X Y	X overlaps with Y
$X \mathbf{s} Y$ $Y \mathbf{si} X$	X Y	X starts Y
$X \mathbf{d} Y$ $Y \mathbf{di} X$	X Y	X during Y
$X \mathbf{f} Y$ $Y \mathbf{fi} X$	X Y	X finishes Y
$X = Y$	X Y	X is equal to Y

Figure 5: (Allen, 1983)'s interval algebra for describing the thirteen possible temporal relation ships between two observed intervals

6 Conclusion

We have presented a corpus study with a novel dialogue taxonomy for motor skill coaching dialogues. We argue this puts requirements on formal models of situated dialogue, including fine-grained shared time representations, and characterizing what is under discussion and in the common ground– in these kind of domains this is generally not questions and propositions, but skills and their desired and observed parameters. In future work we wish to analyze skill referencing completely multimodally, rather than in the verbal sense with accompanying non-verbal acts as we do here[6] and also investigate how the grounding status, both grounded$_{understanding}$ and grounded$_{skill}$, of skills under discussion generalizes to other learning domains.

Acknowledgments

We thank Cornelia Frank for help with the corpus collection, Angelika Maier for annotation and Gerdis Anderson, Michael Bartholdt and Oliver Eickmeyer for transcription. We thank the three SemDial reviewers for their helpful comments. This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

---

[6] Thanks to a reviewer for this suggestion.

References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Kristy Elizabeth Boyer, Mladen A Vouk, and James C Lester. 2007. The influence of learner characteristics on task-oriented tutorial dialogue. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS*, 158:365.

Kristy Elizabeth Boyer, Robert Phillips, Michael Wallis, Mladen Vouk, and James Lester. 2008. Balancing cognitive and motivational scaffolding in tutorial dialogue. In *Intelligent Tutoring Systems*, pages 239–249. Springer.

Kristy Elizabeth Boyer, William J Lahti, Robert Phillips, MD Wallis, Mladen A Vouk, and James C Lester. 2009. An empirically-derived question taxonomy for task-oriented tutorial dialogue. In *Proceedings of the Second Workshop on Question Generation*, pages 9–16.

Herbert H Clark. 1996. *Using language*, volume 1996. Cambridge university press Cambridge.

Mark Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, pages 28–35. Boston, MA.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, 48(4):612–618.

Stefan Kopp, Herwin van Welbergen, Ramin Yaghoubzadeh, and Hendrik Buschmeier. 2014. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces*, 8(1):97–108.

Diane J Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration Papers at HLT-NAACL 2004*, pages 5–8. Association for Computational Linguistics.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: The bielefeld speech and gesture alignment corpus (saga) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Terry McMorris. 2014. *Acquisition and performance of sports skills*. John Wiley & Sons.

M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer. 1995. Disfluency annotation stylebook for the switchboard corpus. ms. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Antoine Raux and Mikio Nakano. 2010. The dynamics of action corrections in situated interaction. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 165–174. Association for Computational Linguistics.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Richard A Schmidt. 1975. A schema theory of discrete motor skill learning. *Psychological review*, 82(4):225.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer.

Defining the Right Frontier in Multi-Party Dialogue

Julie Hunter*, **Nicholas Asher**†, **Eric Kow***, **Jérémy Perret***, **Stergos Afantenos***
*IRIT, Université Paul Sabatier
† IRIT, CNRS
juliehunter@gmail.com, nicholas.asher@irit.fr, eric@erickow.com,
jeremy.perret@irit.fr, stergos.afantenos@irit.fr

Abstract

We develop a Right Frontier Constraint (RFC) for multi-party dialogue ("multilogue"), after arguing that extant definitions of the RFC, and in particular that of SDRT, cannot be directly extended to multilogue. Our proposal is developed and tested on a corpus of chats from an online version of the game *The Settlers of Catan*.

Many theories of discourse structure posit a Right Frontier Constraint (RFC) on discourse attachment (Polanyi and Scha, 1984; Polanyi, 1985; Webber, 1988). The RFC restricts the attachment of newly processed units of a discourse to a small subset of the units in the structure already constructed for some portion of the discourse. The motivating hypothesis behind the RFC is that discourse structure plays a major role in controlling salience. A coherence relation R inferred between two bits of a discourse d will have a particular effect on the shape of the overall tree or graph used to represent d's structure in a way determined by the semantics of R and the discourse theory in use. Relations thus determine what nodes are found along the tree or graph's Right Frontier (RF), a set that evolves dynamically as a discourse proceeds. The RF *constraint* captures the observation that new utterances are normally attached to these nodes, which are predicted to be the most salient.

The RFC constrains semantic phenomena like anaphora and topic, as antecedents for most anaphoric expressions and ellipses are hypothesized to be found along the RF (Polanyi, 1985; Webber, 1988; Asher, 1993). It is also potentially helpful for discourse parsing: restricting attachments to units on the RFC considerably reduces the search space for attachments for discourse units and thus has the potential to improve inter-sentential attachment scores, which are in general much lower than scores for intra-sentential attachment (Joty et al., 2015). Note, however, that the RFC rarely on its own determines attachment, and it can be violated in certain discourse configurations (Asher, 1993), though violations are rare in our corpus study (§4.3). The RFC is a defeasibly necessary but not sufficient constraint.

More importantly, the RFC is practically the only structural constraint on discourse attachment that takes the overall *structure* into account. Most discourse parsing models optimize probabilities for attachments over pairs of elementary discourse units, based on features like textual distance or grammatical or lexical properties of the paired elements. While local features are useful, discourse parsing performance lags behind syntactic parsing, because it does not use global features, in the way syntactic methods have done since (Collins and Duffy, 2002). The RFC is just such a global feature: it says the overall structure of the discourse graph has to have a certain shape. Because of data sparseness and our current limitations to supervised learning, it is infeasible to learn probabilistic global constraints like the RFC from the data directly. So defining an appropriate RFC via symbolic methods is a necessary step to improve discourse parsing.

The RFC has in practice been developed for, and tested on, monologue, generally in the form of newspaper texts (Afantenos and Asher, 2010). It is expected to be helpful as a constraint on multilogue as well, though important differences between multilogue and monologue prevent a trivial extension of standard RFC definitions. In monologue, a speaker is uniquely responsible for the information presented in the discourse, and the RFC is a constraint on the way that information should be presented. In dialogue, we deal not only with how speakers present information but also how they pick up on information presented by others. One speaker might make multiple points, but her respondent might pick up on just one, or ignore them all. Or one or more respondents might wish to discuss multiple points simultaneously, introducing multiple conversation threads.

This paper develops a modified RFC suitable for multilogue and makes precise the RFC as a general constraint on discourse parsing. §1 reviews one version of the RFC for monologue, §2 introduces the corpus that we will use to develop our modified RFC, and §3 explains our choice of theoretical framework. In §4, we first extend the RFC to handle certain phenomena found in our corpus that are independent of multilogue (§4.1), and then extend this modified RFC to one suitable for multilogue (§4.2). §4.3 describes some experimental results with this RFC on our corpus. §5 and §6 present open problems and related work.

1 Modelling the RFC for monologue

In general, when an utterance u is made, the content of the utterance immediately prior to u will be highly salient, but other contents might be salient as well. A speaker might linger on a topic—elaborating on it, providing background on it, or explaining it and so on. In such a case, the point that is being elaborated on or explained, etc. will remain salient, and potentially form a chain of salient and accessible contents underneath it.

On the other hand, when a speaker, say, lists a series of attributes or describes a sequence of events, the most recently described attribute/event will be more salient than the previously described ones, rendering the latter inaccessible to later utterances. Thus in (1), the content of π_1 is inaccessible to that of π_3—we cannot infer the sequence $\pi_1 + \pi_3 + \pi_2$, even though that would yield a more coherent discourse (without further context).[1]

(1) Rose dumped the cookies on the floor.$_{\pi_1}$ (So) She was sent to her room.$_{\pi_2}$ (And) She drew all over the kitchen wall.$_{\pi_3}$

If we reverse the order of π_2 and π_3, as in (2), we can group Rose's two acts together, as desired. What's more, while π'_1 alone is inaccessible to π'_3, the fact that π'_2 clearly describes an event in a series of related events makes the group $\pi'_1 + \pi'_2$ salient and accessible. That is, we understand Rose's being sent to her room as the result of both acts, not just of the more recently described one.

(2) Rose dumped the cookies on the floor.$_{\pi'_1}$ (And) She drew all over the kitchen wall.$_{\pi'_2}$ (So) She was sent to her room.$_{\pi'_3}$

To make this precise, let's consider the RFC as defined in *Segmented Discourse Representation Theory* or SDRT (Asher and Lascarides, 2003). In SDRT, the structure for a discourse d is modelled as a rooted spanning DAG (SDAG), called an SDRS, $G = (V, E_1, E_2, \mathit{Last})$. V is the set of *elementary discourse units* (EDUs; labelled $\pi_0, ..., \pi_n$) and *complex discourse units* (CDUs) in d, where an EDU is a clausal or sub-clausal unit and a CDU is a collection of EDUs (and possibly other CDUs) that together serve as an argument to a discourse relation. $E_1 \subseteq V \times V$ is the set of edges or labelled discourse attachments between elements of V. $E_2 \subseteq V \times V$ is the parenthood relation that relates CDUs to their component DUs. We write $e(\pi_x, \pi_y)$ when e is an edge with initial point π_x and endpoint π_y. *Last* is the last EDU in V, following the linear ordering of EDUs determined by their order in d. An SDRS is "spanning" in that all elements of V other than the root have at least (and possibly more than) one incoming edge: $\forall \pi_x \in V.(\pi_x \neq \text{ROOT} \rightarrow \exists \pi_v \in V.((\pi_v, \pi_x) \in E_1))$.

The set E_1 can contain two types of edges, *coordinating* and *subordinating*. Relations such as Explanation, Elaboration, and Background—in which the second argument extends the discussion about the first—are represented with subordinating (vertical) edges. Relations such as Continuation, Narration, and Result—in which the second argument shuts off the accessibility of the first—are represented with coordinating (horizontal) edges. Suppose we prefix (2) with π_0, *We've been having a rough time*, so that π'_1–π'_3 elaborates on π_0. π_0+(2) would yield the graph $G_{\pi_0+(2)}$:

- $V = \{\pi_0, \pi'_1, \pi'_2, \pi'_3\}$
- $E_1 = \{\langle \pi_0, C_1\rangle, \langle \pi'_1, \pi'_2\rangle, \langle C_0, \pi'_3\rangle\}$
- $E_2 = \{\langle C_0, \pi'_1\rangle, \langle C_0, \pi'_2\rangle, \langle C_1, C_0\rangle, \langle C_1, \pi'_3\rangle\}$
- *Last*= π'_3.

---

[1] Eliciting intuitions about examples like (1) is a delicate matter. While rhetorical theories hold that discourse structure and coherence are intimately related, this does not mean that other factors, such as intonation and word choice, do not affect coherence. In (1), it is important to read the example with a normal intonation. Were a speaker to preface π_3 with *and* and pronounce *and* with a certain intonation, it would be clear that she wanted to retroactively add π_3 to the list of reasons why Rose was sent to her room, i.e. π_3 could attach to π_1. However, the special intonation would arguably be a signal that the speaker wanted to return to a less salient point.

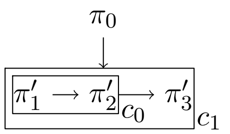


Figure 1: Graph of π_0 + (2)

For monologue, a node π_x is on the RF of a graph G, i.e. $\text{RF}_G(\pi_x)$, just in case π_x is *Last*, π_x is related to *Last* via a series of subordinating (*Sub*) edges, or π_x is a CDU that includes a node in RF_G:

Definition 1 *Let $G = (V, E_1, E_2, Last)$ be a discourse graph. $\forall \pi_x, \pi_y, \pi_z \in V$, $\text{RF}_G(\pi_x)$ iff*
(i) *π_x= Last,* ***(ii)*** *$\text{RF}_G(\pi_y)$ & $\exists e \in E_1$, $e(\pi_x, \pi_y)$ & Sub(e), or* ***(iii)*** *$\text{RF}_G(\pi_y)$ & $\exists e \in E_2$, $e(\pi_x, \pi_y)$.*

So the RF of $G_{\pi_0 + (2)}$ is $\{\pi'_3, C_1, \pi_0\}$. Note that the RF is updated dynamically each time a new EDU is processed; the RF for (attachment of) an EDU π_n will be determined by the graph $G_{\pi_0 - \pi_{n-1}}$. The RF for a CDU $\pi_m \ldots \pi_n$, $m < n$, is the RF for π_m.

2 The *Settlers* Corpus

The Settlers of Catan is a win-lose game in which players trade resources (e.g. wood and sheep) to build roads and settlements. In the standard online version, players interact solely through the game interface, making trades and building roads, etc., without saying a word. In our online version, players were asked to discuss and negotiate their trades via a chat interface before finalizing them non-linguistically via the game interface. As a result, players frequently chatted not only to negotiate trades, but to discuss numerous topics, some unrelated to the task at hand.

The *Settlers* corpus is ideal for studying multilogue. The chats maintain the advantage of written text (no need for transcription) but they manifest phenomena particular to multilogue, such as multiple conversation threads. Also, the chats move quickly, which limits descriptively robust comments and forces players to exploit textual, discourse structuring clues.

The corpus consists of 59 games out of which 36 games (1027 dialogues, 9888 EDUs and 10181 relations) have so far been annotated for discourse structure in the style of SDRT, with a development subset of this corpus containing 9422 relations. This large annotation effort was carried out by 4 annotators who had no special knowledge of linguistics, but who received training over 22 negotiation dialogues with 560 turns. Because annotating full discourse structures is a very complex task (using an exact match criterion of success, the inter annotator agreement score was a Kappa of 0.45 (Afantenos et al., 2012)), experts made several passes over the annotations from the naive annotators, improving the data and debugging it. The 4 naive annotators received no explicit instructions to obey SDRT's RFC, and while expert annotators were aware of the constraint for monologue, they decided collectively not to make attempts to annotate in compliance with it; they picked attachment sites according to their best judgement.

3 Why SDRT?

We have chosen SDRT as the framework to develop an RFC for multilogue. The *Settlers* corpus is already annotated for discourse structure in the style of SDRT and in addition, SDRT's RFC has been empirically validated on written monologue (newspaper articles and Wikipedia entries) using an annotation task in which annotators were not told about the RF, much less instructed to follow it (Afantenos and Asher, 2010). More importantly, however, SDRT deals easily with long distance attachments, which Ginzburg (2012) finds attested in multilogue, and has a semantics capable of dealing with fragments or non sentential utterances (Schlangen, 2003), which are frequent in our corpus. Also, it can model non-tree like structures, like that shown in Figure 2, which account for at least 9% of the links in our corpus. Such structures make theories that model discourse structures with rooted trees, like *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1987) or simple dialogue models where attachments are always made to *Last*, cf. (Schegloff, 2007; Poesio and Traum, 1997), unsuitable. In Figure 2, QAP is the relation Question-Answer-Pair, ACK is Acknowledgement, and "kk" means "okay, cool".[2]

From the perspective of discourse processing, the RFC could be key in solving the *attachment problem*—that of predicting where a discourse unit π_n will attach to the structure for π_0–π_{n-1}. If there are no constraints on a theory of attachment, the search space of solutions is very large making good attachment predictions impossible given the limited amount of data. So adding constraints is potentially crucial. Of course, if attachment is already very constrained, adding an RFC makes little

---

[2] To save space, we skip turns in examples when the turns are irrelevant to our main point.

234	gw	anyone got wheat for a sheep?
235	inca	sorry, not me
236	ccg	nope. you seem to have lots of sheep!
238	dmm	i think i'd rather hang on to my wheat
239	gw	kk I'll take my chances then...

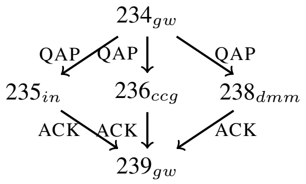


Figure 2: Example of a non-tree-like structure

to no difference. In RST, attachment is restricted to adjunction over trees from contiguous spans, so the attachment problem is comparatively easy to solve; attachment is even more trivial in a theory of dialogue where attachments must be made to *Last*. Such theories would gain little to nothing from an RFC.

SDRT is more liberal in its attachment principles than RST: though it incorporates constraints like connectedness, acyclicity and constraints on CDUs (Venant et al., 2013), non-adjacent and long distance attachments are common. Thus, adding an RFC to SDRT in principle greatly reduces the search space for attachment. When we combine this with the fact that SDRT's graphs can deal with examples like Figure 2 and the examples of multiple threads discussed below, using SDRT to develop an RFC for multilogue is a natural choice.

4 Modifying the RFC

4.1 First modifications

SDRT's RFC relies on an incremental construction procedure that ensures that each EDU π_n is attached at some point along the RF of a connected graph G for EDUs $\pi_1, ..., \pi_{n-1}$ before π_{n+1} is even considered. Before developing an RFC for multilogue, we first need to modify this procedure to handle two phenomena: CDUs and backwards links. This subsection treats these topics in turn.

The incremental construction procedure assumes that it is possible to tell where a CDU will attach to an incoming discourse structure even before the full content of the CDU is known. Given that a CDU is a group of DUs that function together to form a single argument to a discourse relation, the incremental procedure potentially introduces a fair amount of guesswork into the process of reasoning about attachment. Consider (3) and the two possible continuations, (a) and (b).

(3) Bill: I'm running late$_{\pi_0}$ because my car broke down$_{\pi_1}$.
Janet: If you call Mike$_{\pi_2}$, ...
a. he might be able to pick you up and get you to the party on time$_{\pi_3}$.
b. he might be able to come over and fix your car$_{\pi'_3}$.

In (3a), $\pi_2 + \pi_3$ intuitively attaches to π_0, while (3b) suggests an attachment of $\pi_2 + \pi'_3$ to π_1. Until Janet utters the consequent, we can't tell where she is going with the antecedent.

There are two solutions to the problem posed by CDUs without resorting to a probabilistic version (which does not seem automatically learnable): (i) allow graphs to be corrected/repaired in light of new information (Asher, 1993) or (ii) wait to attach CDUs to an incoming discourse until the content of the CDU is complete. As an illustration, consider the graph G, shown in Figure 3. We can, as shown in (i), construct G by first drawing an edge e_1 from π_x to π_y and then adding an edge e_2 from π_y to π_z and correcting e_1 so that its endpoint is the CDU (π_y+π_z). Alternatively, as shown in (ii), we can wait to draw an edge with π_x as initial point until the CDU (π_y+π_z) has been constructed. (Relevant steps are separated by commas.)

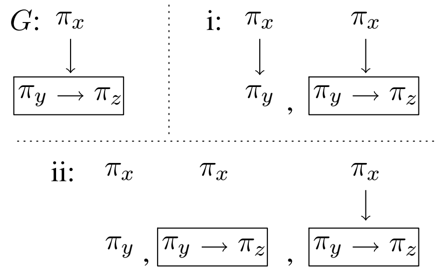
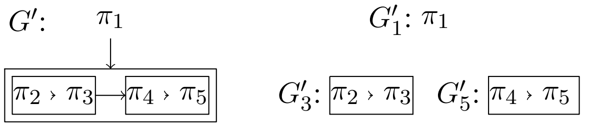

G: π_x
$\pi_y \longrightarrow \pi_z$
i: π_x
π_y ,
π_x
$\pi_y \longrightarrow \pi_z$
ii: π_x
π_y ,
π_x
$\pi_y \longrightarrow \pi_z$,
π_x
$\pi_y \longrightarrow \pi_z$

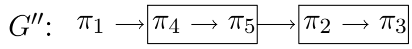
Figure 3: Corrected vs. delayed CDU construction

We adopt option (ii) and recast the RFC as a constraint on attaching subgraphs. This makes the construction of an SDRS more compositional and allows us to wed the RFC with standard, non-incremental discourse parsing models. Even the standard case of EDU attachment can be thought of in this way. Let π_5 be an EDU that needs to be attached to a connected discourse graph $G_1 = \langle\{\pi_1, \pi_2, \pi_3, \pi_4\}, E_1, E_2, \pi_4\rangle$ and treat π_5 as the sole node in a graph $G_2 = \langle\{\pi_5\}, \emptyset, \emptyset, \pi_5\rangle$. The problem of attachment for π_5 can be recast as the problem of attaching G_2 to G_1.

To verify that a graph G contains no RF violations, we must be able to check for any subgraph of G, whether that subgraph violates the constraint. And we must allow that a subgraph of G might contain further, unconnected subgraphs, $G_1, G_2, ...G_n$, each with its own *Last*. Let G be an SDRS over EDUs $\{\pi_1, \ldots \pi_j, \pi_{j+1}, \ldots \pi_k, \pi_{k+1}, \ldots \pi_n\}$ and suppose we have constructed three subgraphs $G_j = G \upharpoonright \{\pi_1, \ldots \pi_j\}$, i.e. G restricted to $\pi_1, \ldots \pi_j$ in their textual order, $G_k = G \upharpoonright \{\pi_{j+1}, \ldots \pi_k\}$, and $G_n = G \upharpoonright \{\pi_{k+1}, \ldots \pi_n\}$. G_j, G_k, and G_n each has its own RF, open to attachment, which makes possible highly undesirable graphs. Consider G' below and its subgraphs G'_1, G'_3, and G'_5:



If we allow any subgraph to attach to the RF of any other subgraph, we could in theory, combine the subgraphs of G' to build a graph G'' as follows:



In fact, every EDU in any graph G could be considered a single-node subgraph, in which case allowing attachment on the RF of any graph would render an RFC pointless. An utterance could provide the output for a link to an arbitrarily later utterance, and speakers would be able to respond to points that haven't been salient for some time.

G'' is problematic because the CDU $\pi_2 + \pi_3$ is attached to $\pi_4 + \pi_5$, but neither π_4, π_5, nor $\pi_4 + \pi_5$ is on the RF for π_2. Moreover, the RF for a new EDU, π_6, would be defined by π_5 (*Last* in G''), despite the the coordinating link from $\pi_4 + \pi_5$ to $\pi_2 + \pi_3$, which should block attachment to π_5.

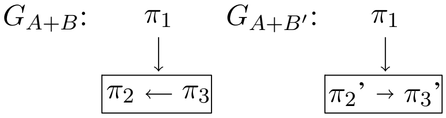
We need to constrain graph development. Let's return to our subgraphs G_j, G_k, and G_n of G, and let G_{jn} be the extension of G_j with G_n. We must eventually construct a graph that attaches G_k to G_{jn}; call it $G_{jn} + G_k$. Such configurations can occur when G_k contains a parenthetical remark about G_{jn} or when it provides the topic. This means that G_k will be subordinate to G_{jn} or that $\text{RF}_{G_k} \cap \text{RF}_{G_{jn}+G_k} \neq \emptyset$. Let $\text{RFC}(G_{jn})$ mean that each edge in G_{jn} complies with the RFC in that each node π_n in G_{jn} attaches to a node on the RF for π_n as defined in Definition 1. The predicate OK, defined below, constrains the construction of graphs like G_{jn}. Note that Axiom 1 requires G_k be non-empty.

Axiom 1 *Let* $G = G_{jn} + G_k$*, with* G_j*,* G_n*,* G_k *and* G_{jn} *as described above. Then* $\text{OK}(G)$ *iff:*
(a) $\text{RFC}(G_{jn}) \wedge \exists e(e(G_{jn}, G_k) \wedge Sub(e))$ *or*
(b) $\exists \pi_x(\text{RF}_{G_k}(\pi_x) \wedge \text{RF}_{G_{jn}+G_k}(\pi_x))$

We apply this axiom below.

Another complication, given that edges in E_1 are directed, is that the direction of some edges reverses the textual order of their arguments.

(4) A [Would anyone give me some clay?]$_{\pi_1}$
B [I would,]$_{\pi_2}$ [if you give me a sheep]$_{\pi_3}$
B' [if you give me a sheep]$_{\pi'_2}$ [I would,]$_{\pi'_3}$



A+B yields a coherent SDRS, yet the backwards link $\pi_2 \leftarrow \pi_3$ violates the RF defined by Definition 1. The EDU π_1 is *Last* from the point of view of π_2, and so defines the RF for π_2; π_3 will not figure in this RF, thus the edge from π_3 to π_2 is a violation.

Furthermore, while (4B) is truth conditionally equivalent to (4B'), they are not discourse equivalent because (π_2 + π_3) and (π'_2 + π'_3) do not have the same felicitous continuations; i.e., $(\pi_x \rightarrow \pi_y)$ and $(\pi_y \leftarrow \pi_x)$ make importantly different contributions to discourse structure.

(5) [I would,]$_{\pi_2}$ [if you give me a sheep.]$_{\pi_3}$
a. [and an ore]$_{\pi_4}$
b. ??[with pleasure.]$_{\pi'_4}$

(6) [if you give me a sheep]$_{\pi'_2}$ [I would.]$_{\pi'_3}$
a. ??[and an ore]$_{\pi_4}$
b. [with pleasure]$_{\pi'_4}$

The examples above are noticeably more felicitous if the continuation targets the textually last EDU (π_3 or π'_3) despite the fact that these EDUs are the inputs for their respective conditional links.

To handle backwards links, we permit two graphs G_n and G_m to be attached with an edge in either direction. $\text{RFC}(G, e(\pi_x, \pi_y))$ means that the edge e complies with the RFC in G. We define an undirected RFC constraint over graphs G_n and G_m of an eventual graph G by extending Definitions 1 Axiom 1 with Axiom 2:

Axiom 2 *$\forall\pi_x{\in}V^{G_n}$, $\forall\pi_y{\in}V^{G_m}$ such that $\neg\exists e \in E_1^{G_n} \cup E_1^{G_m}$. $(e(\pi_x,\pi_y) \vee e(\pi_y,\pi_x))$:*
RFC$(G_n + G_m, e(\pi_x,\pi_y))$ *iff*
(a) RF$_{G_n}(\pi_x) \wedge$ RF$_{G_m}(\pi_y)$ *or*
(b) RF$_{G_n}(\pi_y) \wedge$ RF$_{G_m}(\pi_x)$

The full definition of an undirected RFC, RFC$_u$, over the fusion of any two subgraphs now is:

Definition 2 RFC$_u(G_n{+}G_m)$ *iff* $\forall e \in (E_1^{G_n+G_m} \setminus (E_1^{G_n} {\cup} E_1^{G_m}))$ *:* OK$(G_n{+}G_m) \wedge$ RFC$(G_n{+}G_m, e)$

We can now handle examples (5)-(6). Consider (6). In constructing the graph for (6a), π_2' and π_3' potentially determine separate subgraphs. Suppose we attach π_4 to π_2' to build the structure $[\pi_2' \rightarrow \pi_4] \rightarrow \pi_{3'}$ (a felicitous combination of the EDUs in (6a)). $\pi_{3'}$ is the only node on the RF in the subgraph consisting only of $\pi_{3'}$, so by Axiom 1, it should remain on the RF once we attach it to $\pi_2' + \pi_4$, but this will not be the case, as the RF will be defined by π_4, the *Last* node. Hence we predict that (6a) is unacceptable while (6b) is acceptable. Reversing the links makes no difference; while the highest link is reversed in (5), *Last* is determined by textual order, so *Last* is π_3 not π_2. Thus we cannot attach π_4' to π_2 in (5b) for the same reason that we cannot attach π_4 to π_2' in (6a).

4.2 Extending the modified RFC to multi-party dialogue

Our undirected RFC cannot yet handle structures like that in Figure 2 (as neither 235 nor 236 are on the RF for 239) or examples of "interleaved threads", in which speakers juggle multiple conversations simultaneously. Both types of example are common in our corpus; the example in Figure 4 involves (at least) three interleaved threads.

165	lj	**anyone want sheep for clay?**
166	gw	**got none, sorry :(**
167	gw	so how do people know about the league?
168	wm	**no**
170	lj	i did the trials
174	tk	i know about it from my gf
175	gw	[yeah me too,]$_a$
		[*are you an Informatics student then, lj?*]$_b$
176	tk	did not do the trials
177	wm	**has anyone got wood for me?**
178	gw	[I did them]$_a$ [because a friend did]$_b$
179	gw	**lol william, you cad**
180	gw	**afraid not :(**
181	lj	*no, I'm about to start math*
182	tk	**sry no**
183	gw	**my single wood is precious**
184	wm	**what's a cad?**

Figure 4: Example of interleaved threads

To handle such examples, we assign each speaker s in a multi-party dialogue a textual *Last*, i.e. the textually last EDU that s introduced into the chat. We call the RFC defined with individual speaker *Last*s RFC+MLAST. RFC+MLAST allows the discourse parser to attach turns 235, 236 and 238 in Figure 1 to turn 239 without violations, because for every edge with 239 as its endpoint, its initial point is *Last* for some speaker. For Figure 4, MLAST lets 168 (*no*) attach to 165 as an answer, even though GW has introduced a separate question on a completely different topic that attaches via a coordinating Continuation relation to 165. Similarly, MLAST allows us to attach 175b to LJ's turn in 170 and GW's in 178 to 176 in spite of WM's attempt to start a new bargaining session. Likewise for the attachment of 182 to 177. RFC+MLAST fails, however, to allow the intuitive attachment of 181 to 175b, because GW's *Last* is 180 not 175b (see §5 for discussion). Still, it yields considerable improvement over the modified RFC from §4.1. Table 1 shows the effect of MLAST on RFC violations on the development portion of the *Settlers* corpus. The manually annotated structures obey RFC+MLAST on 95% of the links, while only 83.5% of the links obey the RFC from §4.1.

4.3 Experiments and Results for MLAST

A dynamic calculation of restrictions to the search space for attachments using basic RFC and RFC+MLAST shows that RFC+MLAST has a positive effect on the search space for dialogue parsing in the *Settlers* corpus. As shown in Figure 5, the number of possible attachment points decreases dramatically with RFC+MLAST as the size of the dialogues in the corpus increases.

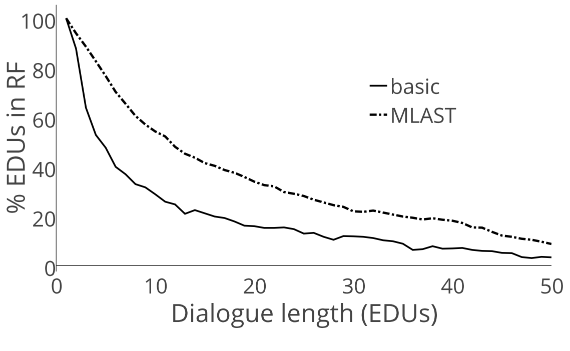


Figure 5: BASIC and MLAST versions of RFC

Using RFC+MLAST can have an important and beneficial effect on parsing. Yet just as the value of adding an RFC can vary depending on the dis-

Data	total links	RFC	MLAST	F-attachment
gold	9293	1536	447	100%
MST	8179	267	191	60.4%
ILP	17430	4342	2693	49.3%
LAST	8179	0	0	56%

Table 1: RFC violations

course theory in question, it can also vary depending on the discourse parser in question. We have developed and trained learner and decoder dialogue parsers for attachment on a simplified version of the *Settlers* chat corpus (without CDUs). The learner is a regularized maximum entropy (MaxEnt) model (Berger et al., 1996). Using standard, superficial features for discourse parsing of the sort found in e.g., Muller et al. (2012) and Li et al. (2014), we learn a probability distribution over pairs of EDUs as an input to several decoders. One decoder uses the MST algorithm (jin Chu and hong Liu, 1965; Edmonds, 1967). Another constructs first a maximal spanning DAG or MSDAG (McDonald and Pereira, 2006; Schluter, 2014) and then prunes it with constraints defined using ILP. The attachment F-scores for MST and ILP without the RFC are provided in Table 1.

As shown in Table 1, MST closely complies with the standard RFC; 96,7% of its predicted attachments obey the RFC while 97,7% comply with RFC+MLAST. Therefore, using RFC+MLAST as a filtering constraint on MST would have little effect. ILP on the other hand could benefit considerably from having RFC+MLAST as a constraint, gaining up to 10% in its attachment score.

The data on MST, however, raise questions about its value as a parsing algorithm for our corpus. Note how closely it complies with the RFC. This is surprising, because CDUs are important in calculating the RF in both monologue and multilogue, so we would expect a considerable amount of RFC violations with a decoder that ignores CDUs. This is especially so given that removing CDUs from the gold annotations on the *Settlers* corpus results in about a 10% increase in violations of the basic RFC; only 73% of the attachments in the manually annotated corpus obey RFC once we drop CDUs.

A baseline where we simply attach each EDU to the preceding one verifies the plain RFC at 100%. We call this baseline LAST. The RFC violations over our corpus suggest that MST is much closer to LAST than it is to the gold annotations. The figures suggest that tree construction algorithms such as MST miss around 12% of the attachments in the gold corpus that are RFC violations but not violations on RFC + MLAST. Thus while MST might be a locally good strategy (with attachment F-scores at 0.81 within a sequence of consecutive turns by the same speaker), it is a globally mediocre strategy. This worsening echoes the difference reported by others between intra-sentential attachment scores and inter-sentential attachment scores in monologue (Joty et al., 2015). ILP, on the other hand, patterns more closely with the gold data and has many more long distance links.

5 Beyond MLAST

Double-tasking Recall that RFC+MLAST blocks the attachment of 181 to 175b in Figure 4, because GW's *Last* is 180, and not 175b. This violation is interesting, because it illustrates a systematic pattern in which the same speaker carries on several interleaved threads, while others are talking. Such cases intuitively call for multiple *Last*s for a single speaker; that is, a *Last* for speaker s for each thread in which s is engaged. This notion, in turn, calls for a criterion for distinguishing threads.

One possible, and simple, solution would be to individuate threads by their members. Then we could extend the RFC+MLAST to include a *Last* for each speaker for each subset of speakers that is engaged in a thread. This would solve the problem of attachment in Figure 4; however, it would not solve the problem in general, as we also have examples of multiple threads involving the very same subset of speakers. In Figure 6, LJ and GW are engaged in both a trade negotiation, which takes place over turns 123-125, 127-129 and 131, and a thread about whether gw took logic, which takes place over turns 119,126 and 130. Even if we add a *Last* for each subgroup of speakers, 126, 128, and 130 will still give rise to RF violations.

119	lj	**gw did you take logic1 this year?**
123	gw	anyone got more clay? I fancy another
124	gw	can offer a range of items
125	lj	i have clay
126	gw	**no i didn't lj, I'm not a student :)**
128	lj	would like wood
129	gw	1 for 1?
130	lj	**ahhh ok, never mind**
131	lj	sure

Figure 6: More interleaved threads in duologue

It is difficult to define a thread precisely. And in fact, it's not clear to us that 126, 128, and 130 shouldn't count as RFC violations, in the same way that "discourse subordinations" (Asher, 1993) in monologue text count as RFC violations. Violations involving multiple threads with the same two speakers can be coherent but they require more effort to understand. For instance, annotators and interpreters could argue about the attachment of 130 to 126; and if we imagine that GW had made a different offer in 129 (say, 2 for 2 or 2 for 1), the we could easily imagine 130 as a response to 129. Moreover, GW actually refers to LJ by name in 126. This is a funny thing to do given that LJ is his only interlocutor at this point; if we treat 126 as an example of discourse subordination, however, then we can imagine that the name is being used as a signal for a discourse subordination.

Turn internal violations While we have not found a significant number of such examples in our corpus, the RFC might ultimately need loosening to handle examples like the following.

(7) B: Who has ore? **I have sheep to give**. *I could also give some clay.*
A': How many sheep?
B': ?? Three sheep even.

(8) A: Anyone want ore for sheep?
B: **I'm not giving up my sheep for now**, *but lj might want to give some of hers.*
A': What if I offer you two ore?
B': ?? Not for all the ore in the world.

Attachment possibilities for speakers are asymmetric. In (7)-(8), the boldface argument is related to the italicized argument by a coordinating relation (Alternation in (7), Contrast in (8)), which should block the accessibility of the boldface argument. Indeed, B cannot continue with a comment targeting this argument (B+B'), though B' *would have been* felicitous in the absence of the italicized argument. By contrast, if another speaker, A, responds to B's turn, both arguments of the coordinating relation are accessible, as shown by the felicity of the A' continuations (B+A').

The theoretical explanation of this has to do with the underlying semantics of contributions in multilogue. The meaning of a dialogue is a set of commitment slates, one for each speaker. Speakers commit to their own contributions in dialogue but not necessarily to the contributions of their interlocutors, *unless* the attachments they make of *their own contributions* requires also that they take on board the commitments of the interlocutor (Hamblin, 1987; Lascarides and Asher, 2009). From this point of view, an asymmetry in the RFC is to be expected in multilogue.

6 Related Work

The RFC is related to projectivity in parsing (Nivre, 2003). Like projectivity, RFC compliance is a property of a graph with respect to textual order, and like projectivity, the RFC rules out crossing dependencies (relative to textual order) except in special cases. Unlike projectivity, however, the RFC depends on a semantic distinction between subordinating and coordinating relations, and a distinction between CDUs and EDUs. Projectivity and the RFC are thus not equivalent even on trees.

The RFC has been a topic of interest in theoretical work on discourse structure for a long time. But to our knowledge, we are the first to study how it fares for multilogue on a large discourse annotated corpus. With regard to empirical work on discourse parsing, Afantenos and Asher (2010) demonstrate the potential of this constraint, but we are not aware of any actual parsing results with the RFC for monologue or dialogue. Afantenos and Asher (2010) also conducted an empirical study on RFC for monologue. However, we have shown that the RFC for monologue is not suitable for multilogue and must be modified.

7 Conclusions

This paper has presented an account of the RFC in multilogue with complex segments, backwards links, and simultaneously running multiple threads of conversation. We have shown our corpus verifies our modified RFC+MLAST. Our experiments have shown that some discourse parsing methods can benefit substantially from the RFC as a processing constraint and that in general the RFC provides an important reduction in the search space of possible attachments. In future work, we will implement our modified RFC for parsing on multilogue data and investigate further the empirical effects of modified LAST to account for the difficulties mentioned in section 5.

Acknowledgements

We thank Semdial reviewers for their comments. This work was supported by ERC grant 269427.

References

Stergos Afantenos and Nicholas Asher. 2010. Testing SDRT's right frontier. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1–9.

Stergos Afantenos, N. Asher, F. Benamara, A. Cadilhac, C. Dégremont, P. Denis, M. Guhe, S. Keizer, A. Lascarides, O. Lemon, P. Muller, S. Paul, V. Popescu, V. Rieser, and L. Vieu. 2012. Modelling strategic conversation: model, annotation design and corpus. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (Seinedial)*, Paris.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: kernels over discrete structures, and the voted perceptron. In Isabelle Pierre, editor, *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 263–270.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71B(233–240).

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Charles Hamblin. 1987. *Imperatives*. Blackwells.

Yoeng jin Chu and Tseng hong Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*. in press.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland, June. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. *International Pragmatics Association Papers in Pragmatics*, 1:79–105.

Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of EACL*, pages 81–88.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India, December. The COLING 2012 Organizing Committee.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*. Citeseer.

Massimo Poesio and David Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.

Livia Polanyi and Remko Scha. 1984. A syntactic approach to discourse semantics. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING84)*, pages 413–419, Stanford.

Livia Polanyi. 1985. A theory of discourse structure and discourse coherence. In P. D. Kroeber W. H. Eilfort and K. L. Peterson, editors, *Papers from the General Session at the 21st Regional Meeting of the Chicago Linguistic Society*. Chicago Linguistic Society.

Emanuel Schegloff. 2007. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge.

David Schlangen. 2003. *A Coherence-based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, University of Edinburgh.

Natalie Schluter. 2014. On maximum spanning dag algorithms for semantic dag parsing. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 61–65, Baltimore, MD, June. Association for Computational Linguistics.

Antoine Venant, Nicholas Asher, Philippe Muller, and Pascal Denis Stergos D. Afantenos. 2013. Expressivity and comparison of models of discourse structure. In *Proceedings of Sigdial 2013*, Metz, France.

Bonnie Webber. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122. Association for Computational Linguistics, Morristown, NJ.

Learning Trade Negotiation Policies in Strategic Conversation

Simon Keizer, Heriberto Cuayáhuitl, Oliver Lemon
Interaction Lab
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh (UK)
`{s.keizer|h.cuayahuitl|o.lemon}@hw.ac.uk`

Abstract

This paper presents a new, data-driven method for learning trade negotiation policies in strategic, non-cooperative dialogue. The learned policies focus on selecting trade offers in the context of playing the game *Settlers of Catan*. First, a supervised learning approach is used to train a Random Forest model for ranking trade offers, using data from an annotated corpus of humans playing the game. Second, a Reinforcement Learning agent for trade offer selection is trained by playing games against three artificial players that use the human-like, data-driven Random Forest offer selection model. In a comparative evaluation our trained models significantly outperform an expert hand-crafted negotiation baseline as well as the supervised learning negotiator. We therefore show that rather than hand-crafting rule-based heuristics for trading, a more successful approach is to train policies from human trading dialogue data.

1 Introduction

Non-cooperative dialogues, where agents act to satisfy their own goals rather than those of other participants, are of practical and theoretical interest (Georgila and Traum, 2011; Efstathiou and Lemon, 2014a). The game-theoretic underpinnings of non-Gricean behaviour have also been investigated (Asher and Lascarides, 2008).

In practice, it may be useful for dialogue agents not to be fully cooperative when trying to gather information from humans, or when trying to persuade, or for believable characters in video games and educational AI (Georgila and Traum, 2011; Shim and Arkin, 2013). Negotiation, where hiding information (and even lying) can be advantageous, is also of interest (Traum, 2008).

Previous work on Reinforcement Learning (RL) in non-cooperative dialogue (Efstathiou and Lemon, 2014a) focused on a small 2-player trading problem with 3 resource types, and without using any real human dialogue data. This work showed that explicit manipulation moves (e.g. "I really need sheep") can be used to win when playing against adversaries who are gullible (i.e. they believe such statements) but also against adversaries who can detect manipulation and can punish the player for being manipulative (Efstathiou and Lemon, 2014b).

In this paper, we apply RL to a much larger trading problem in the context of the board game Settlers of Catan, involving 4 players and 5 resource types to trade with. Furthermore, the trading polices are optimised by playing against adversaries that have been trained on a corpus of trading conversations between humans playing the game.

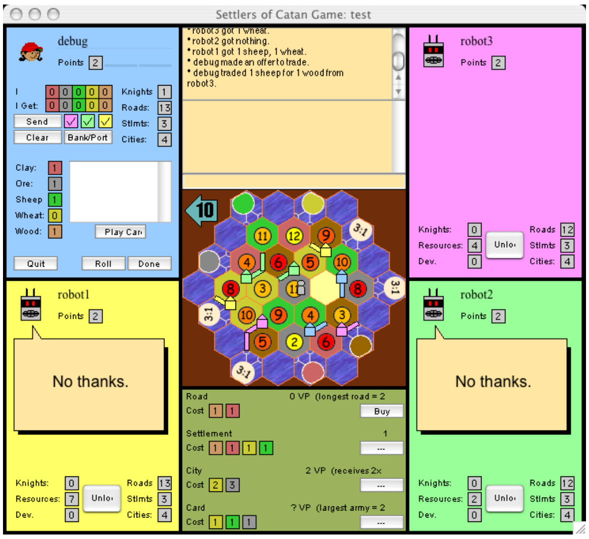
2 Task domain and dialogue data

Settlers of Catan is a complex multi-player board game[1]; the board is a map consisting of hexes of different types: hills, mountains, meadows, fields and forests (see the central part of Fig. 1). The objective of the game is for the players to build roads, settlements and cities on the map, paid for by combinations of resources of different types: clay, ore, sheep, wheat and wood, which are obtained according to the numbers on the hexes adjacent to which a player has a settlement or city after the roll of a pair of dice at each player's turn. In addition, players can negotiate trades with each other in order to obtain the resources they desire. Players can also buy Development Cards, randomly drawn from a stack of different kinds of cards. Players earn Victory Points (VPs) for their settlements (1 VP each) and cities (2 VPs each), and for having the Longest Road (at least 5 consecutive

---

[1] `www.catan.com`

roads; 2 VPs) or the Largest Army (by playing at least 3 Knight development cards; 2 VPs). The first player to have 10 VPs wins the game.



* robot3 got 1 wheat.
* robot2 got nothing.
* robot1 got 1 sheep, 1 wheat.
* debug made an offer to trade.
* debug traded 1 sheep for 1 wood from robot3.

Figure 1: Graphical interface of the online game version of Settlers of Catan, showing in the central area the board itself, in each corner information about one of the four players, and in the top middle area a chat interface with the game history and a text box for chat negotiation.

In our work we are interested in strategic conversation and therefore focus on the trade negotiation aspect of the game. The negotiation models we build take as input a so-called Build Plan (BP), which can be 1) to build a settlement, 2) to build a city, or 3) to buy a development card. Based on the player's current resources and BP, s/he selects trade offers to the other players in order to obtain the resources to realise the BP.

2.1 The jSettlers implementation

For testing and evaluating our models for trade negotiation, we use the jSettlers[2] open source implementation of the game (Thomas, 2003). The environment is a client-server system supporting humans playing against each other via a graphical interface, but also has artificial players. These artificial players use complex heuristics for both the board play (including deciding when and where to build roads, settlements and cities) and negotiation with other players. Our models are integrated in the environment as replacements of the built-in negotiators.

2.2 Human data

With the aim of studying strategic conversations, a corpus of on-line trading chats between humans playing Settlers of Catan was collected (Afantenos et al., 2012b; Afantenos et al., 2012a). The jSettlers implementation of the game was modified to let players use a chat interface to engage in conversations with each other, involving the negotiation of trades in particular. Table 1 shows an annotated trade negotiation chat from the corpus between players W, T, and G; in this dialogue, a trade is agreed between W and G, where W gives G a clay in exchange for an ore.

For the supervised learning experiments, which will be described in Section 3, we used a set of 32 logged and annotated games, corresponding to 2512 trading negotiation events (training instances) denoted as $D=\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i$ are vectors of features and y_i are class labels (i.e. giveable resources). An example trading negotiation in the game of Settlers of Catan in natural language is *"I'll give anyone sheep for clay"*, which can be represented as follows, including the agent's available resources: $Give(Sheep, all) \wedge Receive(Clay, all) \wedge Resources(\,clay = 0,\ ore = 0,\ sheep = 4,\ wheat = 1,\ wood = 0\,) \wedge Buildups(\,roads = 2,\ settlements = 0,\ cities = 0\,)$.

From this example, we extract the training instance $y_i = sheep$ and $\mathbf{x}_i = \{\,clay = 0,\ ore = 0,\ sheep = 4,\ wheat = 1,\ wood = 0, roads = 2, settlements = 0, cities = 0\}$. During each training situation, there is a finite set of possible trading negotiation events. Choosing the best trading offer can be seen as a *ranking task*, where we focus on computing a score representing the importance of each trading offer (similar to the one above)—from which we choose the trade with the highest score. While the model in Section 3 computes the most human-like trading negotiation, the one in Section 4 computes the one with the highest cumulative reward. Other applicable learning approaches are discussed in (Cuayáhuitl et al., 2013).

Note that this corpus was not collected with expert or especially experienced players of the game. We could therefore expect different and more successful trading behaviour to be found in a corpus of expert games.

---

[2] `jsettlers2.sourceforge.net`

Speaker	Utterance	Game act	Surface act	Addressee	Resource
W	*can i get an ore?*	Offer	Request	all	Receivable(ore,1)
T	*nope*	Refusal	Assertion	W	
G	*what for.. :D*	Counteroffer	Question	W	
W	*a wheat?*	Offer	Question	G	Givable(wheat,1)
G	*i have a bounty crop*	Refusal	Assertion	W	
W	*how about a wood then?*	Counteroffer	Question	G	Givable(wood,1)
G	*clay or sheep are my primary desires*	Counteroffer	Request	W	Receivable((clay,?) OR (sheep,?))
W	*alright a clay*	Accept	Assertion	G	Givable(clay,1)
G	*ok!*	Accept	Assertion	W	

Table 1: Example trade negotiation chat.

3 Supervised learning

Here we cast trading in interactive board games as a statistical classification task, where we trained a Random Forest classifier using the features listed in Table 2. Our set of features includes the resources available (features 1-5), the built pieces ('buildups', features 6-8) with a default minimum of 0 and maximum value of 7, the receivable resources in binary form to reduce data sparsity (features 9-13), and the giveable resource contains the classes to predict (feature 14). This agent is trained using an ensemble of trees, which are used to vote for the class prediction at test time (Breiman, 2001; Hastie et al., 2009).

No.	Feature	Domain
1	hasClay	{0...7}
2	hasOre	{0...7}
3	hasSheep	{0...7}
4	hasWheat	{0...7}
5	hasWood	{0...7}
6	hasRoads	{0...7}
7	hasSettlements	{0...7}
8	hasCities	{0...7}
9	recClay	Binary
10	recOre	Binary
11	recSheep	Binary
12	recWheat	Binary
13	recWood	Binary
14	givable	{Clay, Ore, Sheep, Wheat, Wood}

Table 2: Feature set for predicting the offered resource in human-like trades.

We use probabilistic inference to compute the probability of a trade being generated by a human player, given their current resources. The probability distribution—also viewed as a ranking—of a set of trading negotiations is computed as:

$$P(givable|evidence) = \frac{1}{Z}\prod_{t \in T} P_t(givable|evidence),$$

where $givable$ refers to the predicted class, $evidence$ refers to observed features 1-13, $P_t(.|.)$ is the posterior distribution of the t-th tree, and Z is a normalisation constant—see (Criminisi et al., 2012) for further details. Assuming that Y is a set of trades at a particular point in time in the game, extracting the most human-like trade is defined as:

$$y^* = \arg\max_{y \in Y} Pr(\, givable = y \mid evidence \,).$$

Our evaluation metrics include classification accuracy and precision-recall. The former is computed as Accuracy $= \frac{t_p+t_n}{t_p+t_n+f_p+f_n}$ and the latter as F-measure $= \frac{2 \times precision \times recall}{precision + recall}$, where $precision{=}\frac{t_p}{t_p+f_p}$, $recall{=}\frac{t_p}{t_p+f_n}$, t_p=true positives, t_n=true negatives, f_p=false positives, and f_n=false negatives.

Using the data described in Section 2.2 and an ensemble of 100 trees using the features listed in Table 2[3], we obtained a classification accuracy of 65.7% based on a 10-fold cross validation. A break-down of results per prediction class is shown in Table 3. It can be noted that predicting human trades is a difficult task, and that our Random Forest substantially outperforms a majority baseline. This result motivates future work on learning agents with improved predictive power. See (Cuayáhuitl et al., 2015) for further details.

---

[3] The feature set listed in Table 2 reported the best results when compared to other representations—in this paper we only report our best classifier. Other feature sets that we explored include smaller domains (only binary features), larger domains (only non-binary features) smaller and larger sets of features, multiple givables, among others.

Class	Precision	Recall	F-Measure
Clay	0.697	0.686	0.691
Ore	0.644	0.587	0.614
Sheep	0.684	0.751	0.716
Wheat	0.668	0.628	0.648
Wood	0.588	0.613	0.600
All	**0.657**	**0.657**	**0.656**
Majority	0.055	0.234	0.089

Table 3: Classification results of trades from human players in Settlers of Catan.

4 Reinforcement learning

Although some work has been done on using Reinforcement Learning (RL) in the context of Settlers of Catan (Pfeiffer, 2004), this work focused on learning high-level behaviours for playing the entire game. We however focus on the trade negotiation dialogue aspect of the game, which can be seen as strategic conversations in which the participants try to agree on an exchange of resources. In these conversations, each participant has their own personal goals, which inherently conflict with the other participants' goals. Strategies for trade negotiations in Settlers are very complex, given the large state and action space and the fact that the other players' resources are not observable and can only be partially inferred from the trade dialogues that take place. As a first attempt at using RL to optimise such strategies in the full context of the Settlers game, we have designed a Markov Decision Process (MDP) model in which only the given build plan (BP) and the player's own resources make up the observable state, and the actions correspond to 1-for-1 trade offers, addressed to all other players.

At every time step t, an MDP agent perceives the state of the environment $s_t \in S$, selects an action $a_t \in A$, receives a reward signal $r(s_t, a_t) \in \mathbb{R}$ and transitions to a new state $s_{t+1} \in S$. The goal is to find an optimal policy $\pi : S \rightarrow A$, that maximises the cumulative (discounted) reward over time (Sutton and Barto, 1998).

4.1 States and actions

The states of the MDP model are represented via 6 state features (see Table 4): the first feature is the build plan that the negotiator should target, the other features represent how many units of each resource the agent has. The set of possible values for the resource features is restricted by grouping together all quantities of over 3 units. This results in a state space of $3 * 4^5 = 3072$ states. The action space includes all 1-for-1 offers, plus the no-offer action, making $5 * 4 + 1 = 21$ actions (see Table 5). In contrast to the supervised learning negotiator described in Section 3, which re-ranks legal trade offers, the MDP negotiator can offer resources that it does not in fact have, which might be an effective strategy of misleading opponents.

Feature	Values
BuildPlan	{ settlement, city, dev-card }
NumClay	{ 0, 1, 2, at least 3 }
NumOre	{ 0, 1, 2, at least 3 }
NumSheep	{ 0, 1, 2, at least 3 }
NumWheat	{ 0, 1, 2, at least 3 }
NumWood	{ 0, 1, 2, at least 3 }

Table 4: State features for the MDP trade negotiation model.

Action index	Trade offer
0	No offer
1	1 clay for 1 ore
2	1 clay for 1 sheep
⋮	⋮
19	1 wood for 1 sheep
20	1 wood for 1 wheat

Table 5: MDP action space of all 1-for-1 trade offers.

4.2 Reward Function

The reward function that guides the optimisation consists of small penalties for each offer made, a slightly larger penalty if the offer is illegal, a penalty resp. reward for an offer that is rejected by all resp. accepted by at least one of the other players, and a bigger reward for achieving the build plan, e.g. placing a city on the board (see Table 6).

4.3 Adversaries

In order to optimise and test our MDP negotiator agent, we use four different adversaries to play against during training and evaluation. The four

Type of reward	Value
an *offer* was made	-1
an *illegal offer* was made	-3
the offered *trade succeeded*	5
the offered *trade failed*	-5
a *settlement* was built	25
a *city* was built	50
a *development card* was acquired	20

Table 6: Reward function for optimising the MDP trade negotiation strategy.

types (see Table 7) arise from two types of jSettlers 'bots' (i.e. automated players), and two types of negotiators (HEU and SUP). The bots (BOT1 and BOT2) only differ in their building strategy, not their negotiation strategy. The baseline negotiation strategy (HEU) uses heuristics to filter and rank the list of legal trades and then select the top-ranked trade, whereas the supervised learning negotiator (SUP, see Section 3) uses data-driven ranking for trade selection. For more details about game strategies in jSettlers, see (Guhe and Lascarides, 2014).

Adversary	Build strategy	Negotiation strategy
BOT1-HEU	original jSettlers	heuristics baseline
BOT1-SUP	original jSettlers	supervised learning
BOT2-HEU	advanced jSettlers	heuristics baseline
BOT2-SUP	advanced jSettlers	supervised learning

Table 7: Description of negotiation adversaries for training and evaluation.

4.4 Training

Two different MDP policies were trained, both obtained while playing against three copies of one of the two supervised learning adversaries, BOT1-SUP and BOT2-SUP, described above (Settlers is normally a 4-player game). The resulting trained policies are indicated by prefixing the corresponding adversaries with "TRA-". The MDP policies are optimised using Monte Carlo Control (MCC), a basic RL algorithm which processes recorded $(state, action, reward)$ trajectories, updating estimates of the long-term cumulative reward for each state-action pair, stored in the so-called Q-function $Q(s,a) \in \mathbb{R}$. During training, an ϵ-greedy policy is used, i.e., the agent selects a random action with probability $\epsilon = 0.2$ and an action $a' = argmax_a \, Q(s,a)$ otherwise. This setting is used to balance exploitation of the current policy with exploration of the state-action space (Sutton and Barto, 1998).

Figure 2 shows the learning curve for the optimisation of our MDP negotiator playing against 3 copies of the BOT2-SUP adversary negotiator. Each point on the curve represents the result of an evaluation of the MDP over 10k games, played against three copies of the same BOT2-SUP player it was trained on. Figure 3 shows another learning curve, this time in terms of win rates. The 15% win rate at the beginning of training represents a policy that selects random offers (including many illegal ones). After 11k games, the policy reaches the level of 25% that is expected in a 4-player game if the negotiators are identical (indicated by the green horizontal line); after 14k games the policy starts to be stronger than the BOT2-SUP adversaries. The learning (in terms of average reward) seems to converge after about 21k games, but after that still slowly improves. After about 31k games no further improvement is made for another 10k games, so training was ended at 40k games. At that stage, the policy achieves a win rate of 28%.

5 Evaluation

After training the two MDP policies, an evaluation was carried out by letting each of them play 10k games against three copies of each of the four adversaries. Table 8 gives the results in terms of win rates, where each of the columns represents one of the four evaluation conditions.

The same evaluation was also carried out with the adversaries themselves (last two rows of Table 8), making sure that in all evaluations, the build strategy of all four agents was the same and only the negotiation strategy was different for the player being evaluated. Note that in the last two rows, the 'BOT1/2' prefix refers to the building strategy that matches the evaluation condition, i.e. BOT1 when playing against BOT1-HEU or BOT1-SUP, and BOT2 when playing against BOT2-HEU or BOT2-SUP.

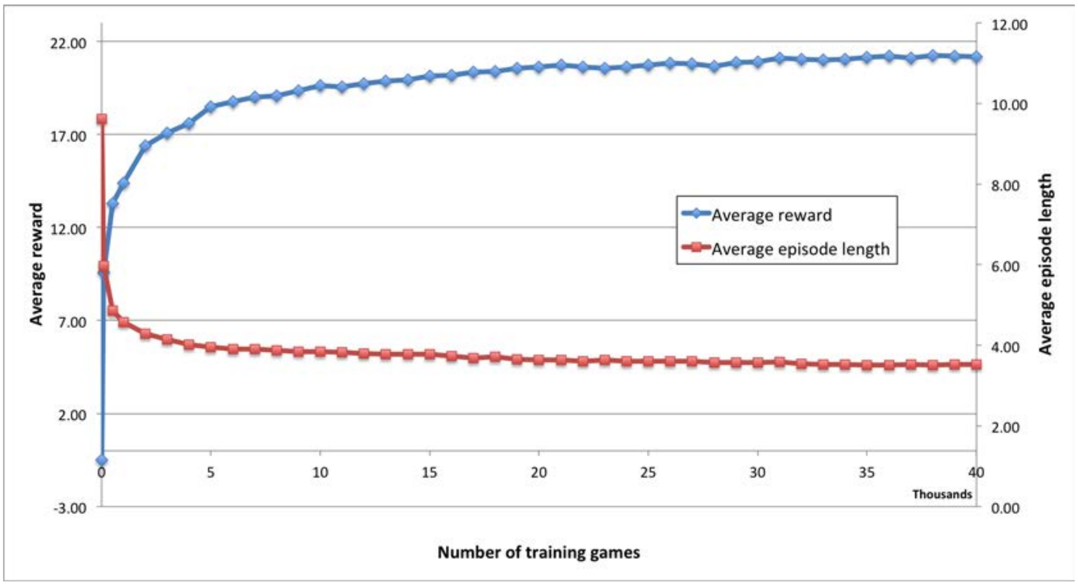


Figure 2: Learning curve in terms of average reward and episode length when training the MDP against the BOT2-SUP adversary.

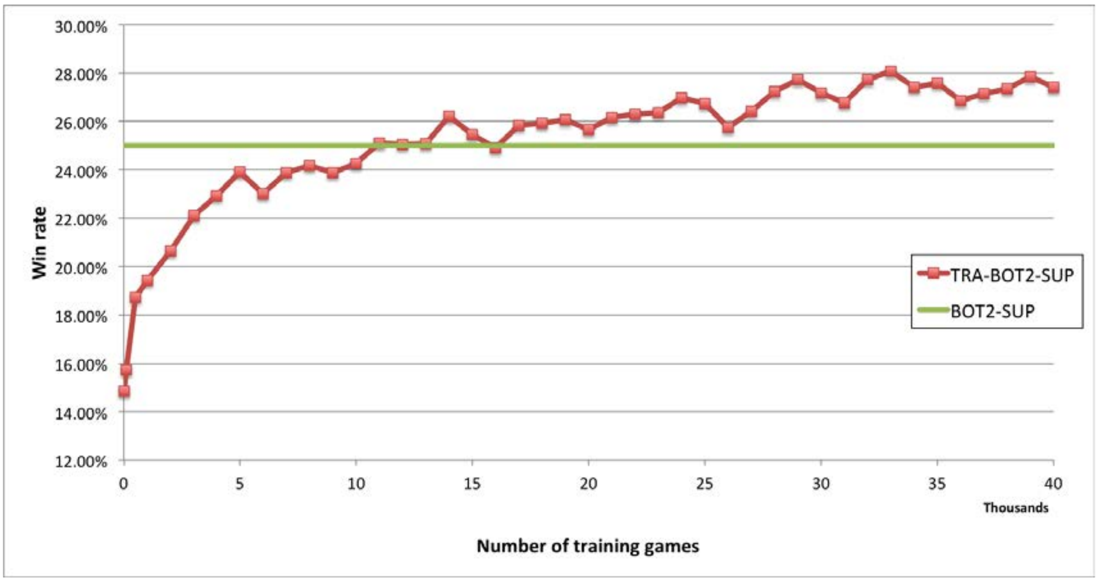


Figure 3: Learning curves in terms of win rates when training the MDP against the BOT2-SUP adversary.

	BOT1-HEU	**BOT2-HEU**	**BOT1-SUP**	**BOT2-SUP**
TRA-BOT1-SUP	**27.35%**	26.41%	**27.26%**	26.23%
TRA-BOT2-SUP	25.90%	**27.69%**	26.81%	**27.40%**
BOT1/2-SUP	24.07%	24.18%	(25%)	(25%)
BOT1/2-HEU	(25%)	(25%)	25.22%	24.99%

Table 8: Evaluation results in terms of win rates of the adversary and trained negotiators (each row corresponding to one negotiator) against the two baseline jSettlers bots (each column corresponding to three instances of an adversary negotiator). In the cases involving four identical players, the expected theoretical win rate of 25% is indicated between brackets.

Note that with 10k games, win rates between 24% and 26% are not statistically significantly different ($p < 0.01$) from the level of 25% that is expected when the four players are identical (Guhe and Lascarides, 2014).

Note also that in the four evaluations of the trained MDP policies, there can be a mismatch between train and test time in the build strategy used: e.g. when evaluating TRA-BOT2-SUP against BOT1-SUP, the trained negotiation strategy is run with the BOT1 build strategy (matching the three adversaries), but it was trained using the BOT2 build strategy. Finally, note that there can also be a mismatch between train and test time in the negotiation strategies of the adversaries.

In all four evaluations, the trained policies significantly outperform both the supervised learning and heuristic baselines. As expected, the policies where the build strategy was the same during training and evaluation obtain higher win rates, as do the ones where the adversary negotiation strategies were the same during training and evaluation. Even when either the build strategy or the adversary negotiation strategies were different between train and test time, the resulting win rates are significantly higher than those of the baselines, except for TRA-BOT2-SUP evaluated against BOT1-HEU, which gets a non-significant 25.9%.

5.1 Discussion

In our experiments we have shown that an MDP model for selecting trade offers can be optimised by interacting with artificial adversaries, and that the optimised model outperforms several baseline strategies in different conditions (i.e. when playing against opponents with different kinds of negotiation strategies).

A preliminary analysis of the trained policies showed that most of the time, trades are selected that one would intuitively expect, given the build plan and resources (e.g. the MDP agent will ask for a resource it needs to realise the given build plan). However, for a portion of the state space the agent seems to select trades that are less obvious, but must have a more indirect long-term game advantage in mind. This can be due to the fact that the negotiation dialogues are embedded in the context of complete games and are therefore not completely isolated from each other. Also, in a number of cases, the trained MDP agent offers resources it does not in fact have, see Table 9. For roughly a third of the states encountered during training, the policy selects an illegal trade offer (column 'PolicyIllegal'), suggesting that lying to the supervised learning adversaries can be advantageous, despite the penalty for illegal offers in the reward function (Table 6). However, many of these states occur quite rarely (making its Q-value estimates relatively inaccurate): in a test run of 1000 games, less than 2% of the turns where the policy is triggered, an illegal offer was selected (column 'IllegalOffers'). After re-evaluating the trained policies whilst restricting the output to legal offers only, no significant differences in terms of win rate were observed (27.75% resp. 27.94% for TRA-BOT1-SUP resp. TRA-BOT2-SUP, tested against adversaries matching the training conditions).

The current MDP model is relatively simple, in the sense that it only supports the selection of 1-for-1 trades, it only takes into account three of the possible build plans in jSettlers, and only knows about its own resources. We therefore plan extensions to both state and action spaces to further improve performance, e.g. by including the option to offer 2 units in the selection of trades, and including strategies for the additional build plans of obtaining the longest road and the largest army (each worth 2 VPs). Since the values of the different components of the reward function used in this paper are qualitatively intuitive but quantitatively somewhat arbitrary, further improvements might be achieved by using data from expert players to *learn* the reward function (Abbeel and Ng, 2004).

A more challenging future direction is to include information about the adversaries' resources and preferences, which are unobservable, but might be estimated based on the trades taking place and the adversaries' negotiation behaviour. This would require a POMDP-style approach (Kaelbling et al., 1998), which attempts to represent and exploit uncertain information. One step further is 'opponent modelling', in which the beliefs, goals, and preferences of the adversaries are modelled and exploited (Gmytrasiewicz and Doshi, 2005), which could be particularly important in non-cooperative dialogue modelling. We also aim to extend the range of actions with strategic conversational moves, for example to manipulate and exploit the adversaries' beliefs and behaviour, following (Efstathiou and Lemon, 2014a; Efstathiou and Lemon, 2014b).

	—POLICY—		—TEST 1000 GAMES—	
	StatesVisited	**PolicyIllegal**	**Turns**	**IllegalOffers**
TRA-BOT1-SUP	2593	29.1%	13,159	1.57%
TRA-BOT2-SUP	2288	32.5%	15,852	1.94%

Table 9: Illegal offer statistics for the learned policies. Column 'StatesVisited' gives the number of unique states encountered during training, column 'PolicyIllegal' gives the percentage of these states for which the policy outputs an illegal offer, and, based on a test run of 1000 games with one of the trained players against against three instances of the adversary matching the training conditions, columns 'Turns' and 'IllegalOffers' give the number of turns and in how many of them the policy selected an illegal offer.

6 Conclusion and future work

In this paper we have presented a new, data-driven method for learning trading negotiation policies in strategic, non-cooperative dialogue. We showed that rather than hand-crafting rule-based heuristics for trading, a more successful approach is to train policies from human trading dialogue data.

The experiments were carried out in the context of the board game *Settlers of Catan*, making use of an existing java implementation of the game, *jSettlers*. First, a supervised learning approach is used to build a Random Forest model for ranking legal trade offers, using an annotated corpus of data from non-expert humans playing the game. Evaluation results for the underlying classification model show significant improvements in terms of accuracy over a majority baseline (see Table 3). Second, an MDP model for trade offer selection is trained using Reinforcement Learning, by playing games against three instances of one of two possible adversaries, which only differ in their build strategy, but both use the Random Forest model for selecting trade offers. In contrast to the adversaries which always select legal trade offers, the MDP model is capable of offering resources that it does not in fact have (i.e. it can lie).

We evaluated our trained MDP policies, again by playing games against three instances of different adversaries. There are four types of evaluation conditions, arising from the adversaries' two possible build strategies and two possible negotiation strategies (one of them being the Random Forest negotiator). We assessed the performance of the policies by comparing their win rates with other player bots playing games under the same conditions. The results indicate that the trained MDPs achieved significantly higher win rates compared to the expected level of 25% that an agent would achieve if it was identical to the three opponents. More specifically, the policies outperform the two baseline negotiation strategies (an expert handcrafted jSettlers strategy and the human-like supervised learning strategy) in terms of win rate when playing against the same three adversaries. The trained MDP is robust in the sense that even when there was a mismatch between train and test time in the build strategy used or the adversary negotiation strategy, the MDP significantly outperforms the baselines in most cases (see Table 8).

Further improvements can be expected when extending both state and action spaces. For example, the current model only supports one-for-one trades, whereas the option to offer more units could make trades more likely to be successful. In addition, improvements in the supervised learning adversary models can also lead to better MDP performance. We also aim to include information about the opponents into the model, e.g. their resources, requiring a POMDP-style approach, since such information is not observable. Opponent modelling would also enable the selection of additional moves for strategic conversation (in addition to lying), such as manipulation moves. Furthermore, we plan to include the selection of other kinds of actions in trade negotiation dialogue that are not yet supported, such as responses to offers (accept, reject, or counter-offer) in our data-driven models. Finally, we aim to eventually test our trained negotiation strategies against humans, which requires modules for Natural Language Generation and Understanding (NLG and NLU), creating an end-to-end text-based dialogue system for playing Settlers of Catan.

Acknowledgments

The research leading to this work is supported by ERC grant 269427 (the STAC project).

References

Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning.

S. Afantenos, N. Asher, F. Benamara, A. Cadilhac, C. Dégremont, P. Denis, M. Guhe, S. Keizer, A. Lascarides, O. Lemon, P. Muller, S. Paul, V. Popescu, V. Rieser, and L. Vieu. 2012a. Developing a corpus of strategic conversation in the settlers of catan. In *Proc. 1st Workshop on Games and NLP*.

S. Afantenos, N. Asher, F. Benamara, A. Cadilhac, C Dégremont, P Denis, M Guhe, S. Keizer, A. Lascarides, O. Lemon, P. Muller, S. Paul, V. Popescu, V. Rieser, and L. Vieu. 2012b. Modelling strategic conversation: model, annotation design and corpus. In *Proc. SEMDIAL*.

N. Asher and A. Lascarides. 2008. Commitments, beliefs and intentions in dialogue. In *Proc. SEMDIAL*.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. 2012. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227.

Heriberto Cuayáhuitl, Martijn van Otterlo, Nina Dethlefs, and Lutz Frommberger. 2013. Machine learning for interactive systems and robots: A brief introduction. In *Proc. MLIS*.

Heriberto Cuayáhuitl, Simon Keizer, and Oliver Lemon. 2015. Learning to trade in strategic board games. In *IJCAI Workshop on Computer Games (IJCAI-CGW)*.

Ioannis Efstathiou and Oliver Lemon. 2014a. Learning non-cooperative dialogue behaviours. In *Proc. SIGDIAL*.

Ioannis Efstathiou and Oliver Lemon. 2014b. Learning to manage risk in non-cooperative dialogues. In *Proc. SEMDIAL*.

Kallirroi Georgila and David Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Proc. of INTERSPEECH*.

Piotr J Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *J. Artif. Intell. Res.(JAIR)*, 24:49–79.

Markus Guhe and A Lascarides. 2014. Game strategies for The Settlers of Catan. In *Proc. IEEE Conference on Computational Intelligence and Games (CIG)*.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134.

M. Pfeiffer. 2004. Reinforcement learning of strategies for settlers of catan. In *Proc. Int. Conf. on Computer Games: Artificial Intelligence, Design and Education*.

J. Shim and R.C. Arkin. 2013. A Taxonomy of Robot Deception and its Benefits in HRI. In *Proc. IEEE Systems, Man, and Cybernetics Conference*.

R. Sutton and A. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

Robert Shaun Thomas. 2003. *Real-time decision making for adversarial environments using a plan-based heuristic*. Ph.D. thesis, Northwestern University.

David Traum. 2008. Extended abstract: Computational models of non-cooperative dialogue. In *Proc. of SIGdial Workshop on Discourse and Dialogue*.

Reducing the Cost of Dialogue System Training and Evaluation with Online, Crowd-Sourced Dialogue Data Collection

Ramesh Manuvinakurike[1], **Maike Paetzel**[1,2] **and David DeVault**[1]
[1]USC Institute for Creative Technologies, Playa Vista, CA, USA
[2]University of Hamburg, Hamburg, Germany
`{manuvinakurike,devault}@ict.usc.edu, 8paetzel@informatik.uni-hamburg.de`

Abstract

This paper presents and analyzes an approach to crowd-sourced spoken dialogue data collection. Our approach enables low cost collection of browser-based spoken dialogue interactions between two remote human participants (human-human condition) as well as one remote human participant and an automated dialogue system (human-agent condition). We present a case study in which 200 remote participants were recruited to participate in a fast-paced image matching game, and which included both human-human and human-agent conditions. We discuss several technical challenges encountered in achieving this crowd-sourced data collection, and analyze the costs in time and money of carrying out the study. Our results suggest the potential of crowd-sourced spoken dialogue data to lower costs and facilitate a range of research in dialogue modeling, dialogue system design, and system evaluation.

1 Introduction and Motivation

The work reported in this paper helps address a critical bottleneck in the design and evaluation of spoken dialogue systems: the availability and cost of collecting human dialogue data for a new domain. When designing, training, or testing a new dialogue system, the collection of in-domain dialogue data, either between two human roleplayers (human-human) or between a human user and a system prototype (human-agent), is both important and expensive. In-domain dialogue data is important because it provides examples of domain-specific language and interaction that serve to highlight important semantic and pragmatic phenomena in the domain, inform system design choices, and also serve as initial training data for system components such as speech recognition, language understanding, and language generation (Lasecki et al., 2013).

At the same time, the collection of this data can be expensive in terms of both time and money. Potential costs include the time needed to locate and recruit participants, the staffing overhead to schedule and coordinate visits by participants to a lab or system installation, and the payment of participation fees. As an example, for the dialogue game discussed in this paper, participants in a recent lab study were paid $15 each, and required the close supervision of a lab staff member for approximately 35 minutes per participant. These costs are substantial, especially when large amounts of data are desired for training system models based on machine learning.

Deploying dialogue systems on the web, and using crowd-sourcing to recruit remote participants, offers the possibility of increasing the availability of participants while simultaneously driving down the costs of data acquisition.

In this paper, we report on a case study in which web-based crowd-sourcing was used to carry out a substantial data collection and evaluation involving 200 remote human participants who played a fast-paced, browser-based image matching game called RDG-Image (Paetzel et al., 2014). The study included 150 participants in human-agent conditions and 50 participants in human-human conditions. By providing a substantial number of human participants at relatively low cost, the study enabled six different system versions to be compared with each other as well as to human-human teams as a baseline.

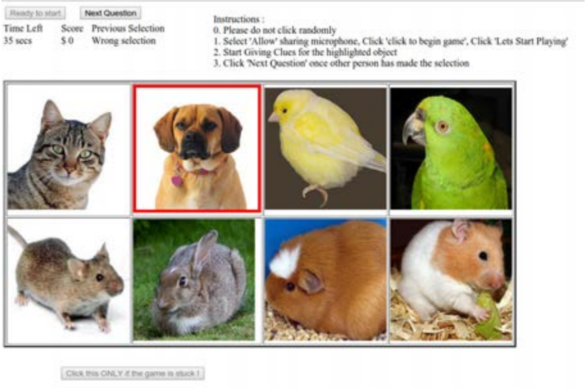
The contributions of the paper are as follows. First, we present and describe how a web-based framework for spoken dialogue data collection, called Pair Me Up (Manuvinakurike and DeVault, 2015), allows for the collection of human-agent spoken dialogues with remote participants. This framework had previously only been applied to human-human data collection. To our knowledge, this is the only current software framework in use by dialogue researchers that can crowd-source both human-human and human-agent dialogue data from remote web users. Second, we report and analyze our case study data collection involving 200 crowd-sourced participants. We discuss the technical challenges we encountered in achieving this data collection, and highlight issues and lessons likely to be valuable to other dialogue researchers who aim to carry out similar crowd-sourced data collections. Finally, we analyze the costs in time and money of carrying out this study, and compare them to the corresponding costs associated with another similar in-lab human-human data collection.

The focus of this paper is on the research methodology of crowd-sourcing as a method of acquiring spoken dialogue data for system development and evaluation. The detailed technical design of our agent and an evaluation of its performance are presented in Paetzel et al. (2015).

We begin in Section 2 with a discussion of the RDG-Image game, which serves as the domain for this study. Section 3 discusses related work on crowd-sourced dialogue data collection. Section 4 briefly summarizes the automated agent used in this study. Section 5 presents our data collection process. Section 6 discusses technical challenges we encountered, and Section 7 presents our analysis of the costs of carrying out this study.

2 The RDG-Image Game

The RDG-Image game is a two player, dialogue-based image matching game (Paetzel et al., 2014; Manuvinakurike and DeVault, 2015). In the game, pictured in Figures 1 and 2, one person plays the role of *director* and the other is the *matcher*. Players are presented a set of eight images. The set of images is exactly the same for both players, but they are arranged in a different order on the screen. One of the images is randomly selected as a target image (TI) and it is highlighted on the director's screen with a thick red border as shown in Figure 1. The goal of the director is to give verbal clues for the TI so that the matcher is able to uniquely identify it from the distractors. Different categories are used for the image sets including pets (Figure 1), fruits, sign language (Figure 2), robots, and necklaces, among others. When the matcher believes he has correctly identified the TI, he clicks on the image and communicates this to the director who has to press a button to continue with the next TI. The team scores a point for each correct guess, with a goal to complete as many images as possible within the stipulated time for each round. Participants are incentivized to score quickly with a bonus of $0.02 per point scored. The player roles remain the same throughout the game. An example of human-human dialogue for a TI is given in Figure 2.



Time Left 35 secs

Score $ 0

Previous Selection Wrong selection

Instructions :
0. Please do not click randomly
1. Select 'Allow' sharing microphone, Click 'click to begin game', Click 'Lets Start Playing'
2. Start Giving Clues for the highlighted object
3. Click 'Next Question' once other person has made the selection

Figure 1: Web browser interface for the RDG-Image game (director's view).

3 Background and Related Work

3.1 Prior Work on Pair Me Up

This study was carried out using a software framework for web-based spoken dialogue collection called Pair Me Up (PMU) (Manuvinakurike and DeVault, 2015). The PMU framework has previously been applied to human-human data collection for the RDG-Image game, and the resulting crowd-sourced data has been analyzed in terms of audio quality, the effect of communication latency, the ability to synchronize collected audio and game events, and the perceived naturalness of remote human-human interactions (Manuvinakurike and DeVault, 2015).

The PMU architecture for human-human data collection is shown in the Figure 3. The system pairs two web users together and connects them into a shared game session where they can converse freely and interact through their browsers. PMU leverages recent developments in web technologies that support development of web-based dialogue systems. It shares this approach with recent dialogue system research such as Jiang et

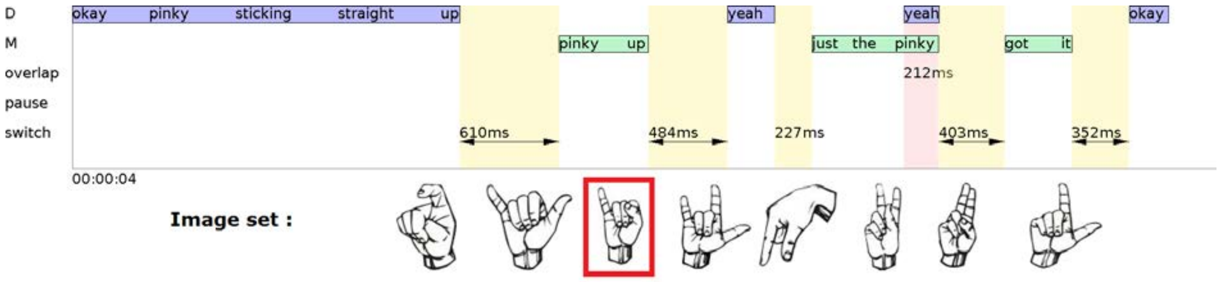


Figure 2: An example from RDG-Image: director D describes the highlighted image to matcher M.

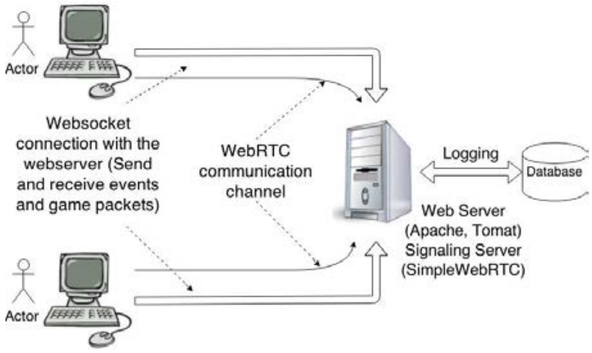

Actor
Websocket connection with the webserver (Send and receive events and game packets)
WebRTC communication channel
Logging
Database
Web Server (Apache, Tomat) Signaling Server (SimpleWebRTC)
Actor

Figure 3: Pair Me Up architecture in human-human mode

al. (2014), which makes use of emerging web technologies to enable a spoken interaction between an individual remote web user and an automated dialogue system. In PMU, several of these new web technologies are used to build an interactive game where the servers can initiate events on remote client browsers, audio is streamed between two remote client browsers, and audio is captured to a server database. Two core technologies the system makes use of are websockets and webRTC. Websockets enable two way communication between the client and server, and they specifically enable the server to push events such as image set changes to the clients, and the clients to send events such as button clicks to the server, without loading a separate URL. The streaming audio communication between the remote clients uses a separate SimpleWebRTC (http://simplewebrtc.com/) channel.

3.2 Prior Work on Crowd-Sourced Dialogue Data Collection

Several large technology companies have recently deployed spoken dialogue systems reaching millions of users on mobile devices (Apple Siri, Google Now, Microsoft Cortana). Such wide deployment suggests the potential in principle for dialogue system builders to acquire large data sets to support designing, training, and evaluating their systems. In the dialogue research community, several researchers have recently taken steps toward collecting dialogue data from systems deployed on the web. Jiang et al. (2014) describe an architecture for capturing typed dialogue interactions in a human-agent configuration, with user speech optionally recognized by Google's cloud-based ASR service. Meena et al. (2014) have also been attracted to crowd-sourcing as a potential source of data, and reported a small-scale experiment in this direction. Some research applications such as Let's Go (Raux et al., 2005) as well as commercial applications (Suendermann et al., 2011; Pieraccini et al., 2009) have collected telephone-based dialogue data from large user populations. One way our work is different from this related work is that our architecture is able to collect both human-human and human-agent spoken dialogues from remote web users.

4 Summary of the agent's design

In this section, we describe the use of the PMU framework for human-agent data collection, briefly summarize the internal design of the agent, and discuss six agent versions used in the study.

4.1 Pair Me Up for human-agent data

The human-agent mode for PMU is configured in a similar way to the human-human mode, as shown in Figure 4. The user connects to the PMU server by following a URL in their browser. A websocket connection is used to transmit game events and system audio between the remote user and the PMU server. The PMU server runs both a webserver process and the automated agent, and these two communicate with each other through TCP sockets. Some modifications were required in PMU to accommodate the

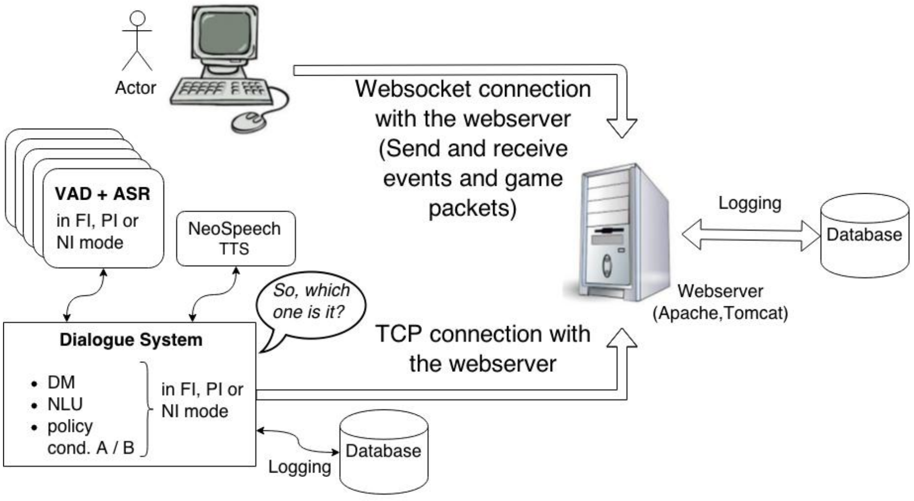


Figure 4: Pair Me Up architecture in human-agent mode

human-agent mode. In human-human mode, bidirectional audio streaming was done through SimpleWebRTC. In human-agent mode, client audio is streamed to the server using HTTP POST requests, and system audio is sent to the client using the websocket.

The agent includes internal modules for Natural Language Understanding (NLU), Dialogue Management (DM), and Dialogue Policy. The agent communicates using TCP socket connections to external processes for Voice Activity Detection (VAD), Automatic Speech Recognition (ASR), Text-To-Speech (TTS), and a database for logging.

4.2 Agent internal architecture

One main design goal for the agent architecture was to build a system which enables us to collect a large data set for multiple agent versions in a short time. On Amazon Mechanical Turk (AMT), there are certain times of the day when many people are available to participate in a study, while during work or sleep hours, among others, data collection is much slower. The peak times can be used best by enabling multiple user interactions at the same time. Thus, we designed the agent such that it can play with multiple users simultaneously, while still keeping track of the dialogue and game states for each user individually. Most agent modules operate in separate threads. This design ensures that the agent is always listening to the user speech, transforming the audio to text, taking decisions and communicating with the dialogue partner at the same time; the use of multiple threads was important to enable the agent to potentially handle multiple users simultaneously with minimal latency. We now briefly summarize the various modules in the agent; see (Paetzel et al., 2015) for additional details.

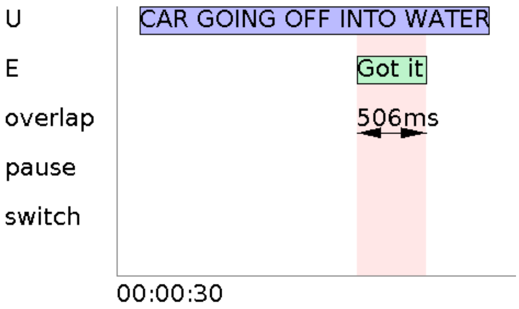


Figure 5: An example from this study: user U describes an image to agent Eve (E).

VAD. Streaming audio from the user's browser is first processed by a Voice Activity Detector (VAD). Detected speech is sent to the ASR either every 100ms or at the end of each VAD segment, depending on the incrementality type (see Section 4.3).

ASR. We use a version of the Kaldi ASR system which is based on (Plátek and Jurčíček, 2014) and was specifically adapted for this study. The ASR provides support for both incremental and non-incremental speech recognition (see Section 4.3).

As audio is streamed into the VAD and ASR, the VAD and ASR both maintain an internal state for decoding the current speech segment. This means one instance of the VAD and ASR cannot serve multiple users at the same time. Thus, multiple instances of the VAD and ASR were running at the same time, with each of them listening to a separate port, as illustrated in Figure 4. The agent takes care of the mapping between a specific user and the respective VAD+ASR instance.

NLU. For language understanding, the agent uses a data-driven statistical classifier to map either partial or final ASR results to one of the eight candidate images on the screen.

DM and Policy. In this study, the agent is always in the matcher role, and its dialogue policy uses statistically optimized rules to decide when the agent should commit to its best guess about the image being described by the user (by saying *Got it!*).

An example of the agent's gameplay is shown in Figure 5. In this example, a picture of a road-sign that warns of a hazardous driving condition is being described.

4.3 Six agent versions

The research motivation for this study is an investigation into the value of alternative types of incremental processing and incremental policy optimization in a system. To support this research, we wanted to run a data collection and evaluation involving six different versions of the agent. While other researchers might not share our specific interest in these six versions, the desire to compare several alternative system designs in an empirical way, ideally using interactive human-agent data, is common to many research efforts.

In our case, our study was designed to evaluate three versions of incrementality and two different policy optimization metrics against each other. The three incremental versions consist of the fully incremental (FI), partially incremental (PI) and non-incremental (NI) versions. Figure 6 illustrates the different versions and their modes of operation. In the FI architecture, the ASR, NLU, DM, and Policy are all operating incrementally after every additional 100 ms of user speech. This setup enables the agent to give fast-paced feedback while the dialogue partner is still talking. For the PI version, only the ASR is operating incrementally; the NLU, DM, and Policy wait for a VAD segment (inter-pausal unit) to finish before they start processing. Here, the agent cannot interrupt the user, but is still able to give a quick response once a pause is detected. In the NI architecture, the ASR, NLU, DM, and Policy are all operating on complete VAD segments as input, which increases the delay between the end of the user's speech and the beginning of the agent's response.

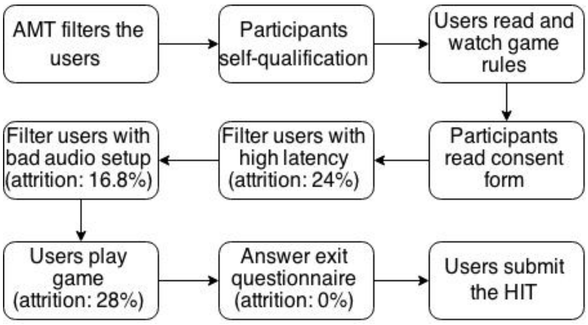

AMT filters the users
Participants self-qualification
Users read and watch game rules
Participants read consent form
Filter users with high latency (attrition: 24%)
Filter users with bad audio setup (attrition: 16.8%)
Users play game (attrition: 28%)
Answer exit questionnaire (attrition: 0%)
Users submit the HIT

Figure 7: The HIT process during the study

Additionally, we optimized policies using two different optimization metrics, which we denote simply A and B in Figure 4. The details of the two optimization metrics are omitted; their technical rationale and motivation is beyond the scope of this paper. Together, the incrementality type and policy type variations creates a 3x2 study design, for a total of six agent versions to evaluate.

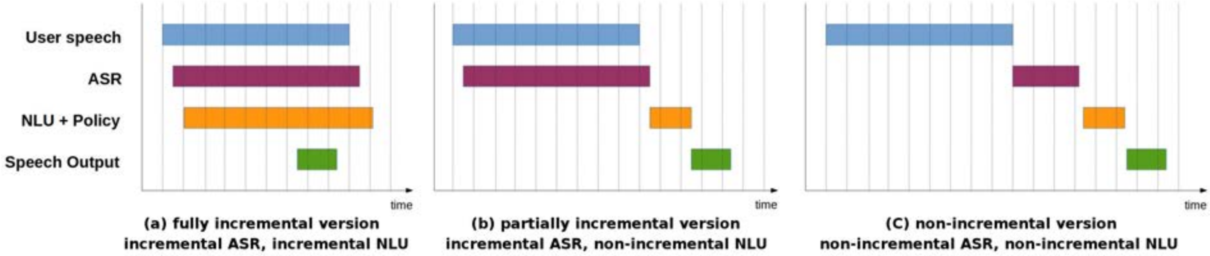
An ability to evaluate so many different agent prototypes empirically is valuable for many research questions, but it also confronts researchers with the difficulty of evaluating them with a significant number of participants in a tight timeframe and with limited financial resources.

The agent's internal modules are designed so that the agent can easily switch between different policies and incrementality types at run-time. Different versions can even be used simultaneously.

5 Crowd-Sourced Data Collection

200 native English speakers aged over 18 were recruited on Amazon Mechanical Turk (AMT) to participate in the study. 25 of them were paired with another human (25 × 2), and 25 played with each of the six versions of the agent (25 × 6). The study was conducted over a period of 10 days. Table 1 summarizes the participant demographics in the study. The study was conducted entirely over the Internet. The protocol involved in recruiting and filtering the participants to guarantee congenial data for the human-agent condition is shown in Figure 7 and discussed in the rest of this section.

AMT filters the users: AMT is able to apply certain filtering criteria for the participants. We had AMT apply the following criteria: (i) Participants have an acceptance rate equal to or greater 92% in their previous Human Intelligence Task



(a) fully incremental version
incremental ASR, incremental NLU

(b) partially incremental version
incremental ASR, non-incremental NLU

(C) non-incremental version
non-incremental ASR, non-incremental NLU

Figure 6: Three different incrementality types in our agent

(HIT) participations; (ii) previous participation in at least 50 HITs; (iii) physical location in the United States or Canada.

Participant's self qualification: The users who AMT qualified for the HIT were provided instructions to participate only if they met the following criteria: (i) must have the latest Google Chrome web browser; (ii) must be a native English speaker; (iii) must have a microphone; (iv) must not be on a mobile device; (v) must have a high speed Internet connection (5 mbps download, 2 mbps upload). Additionally, users were asked to use earbuds or headphones rather than external speakers, which helps prevent their microphone picking up sounds from their own speakers. Next, **users read and watch game rules** in text and video format. The users are then led to a consent form web page where **participants read consent form** and decide if they want to participate or not. The users enter their ID and submit their consent. To prevent certain users with problematic network latency from participating, we measure the network latency between the user and our server (see Section 6.1). 24% of the users who consented to the experiment were filtered out due to high latency or highly variable latency.

Filter users with bad audio setup: The users in the next step were made to listen to an audio file and transcribe it. If the transcriptions were wrong, the users were disqualified. This is to make sure that the users had a functioning speaker/headphone set up. The users then had to speak three pre-selected sentences in their microphone. An ASR transcribed the spoken audio and if the user had at least one word right from the sentences, the users were qualified, else disqualified. 16.8% of the users got disqualified at this step due to a "bad audio set up".

The qualified **users play the game** with the agent. 28% of the users who qualified from the previous stage did not finish playing the game with the agent. It happened that sometimes turkers closed the browser or otherwise stopped participating for reasons we could not discern. After the game, the users were made to **answer an exit questionnaire**. After answering the questionnaire the users were instructed to return to AMT and asked to **submit the HIT**.

	Agent	Human
N	150	25
Female(%)	54.7	44
Age(yrs)		
Mean	31.12	31.12
Median	28	28
SD	10.2	10.4

Table 1: Demographic data for the 175 human directors, based on whether the matcher was an agent or another Human.

6 Technical Challenges Encountered

We faced several technical challenges in achieving this data collection. The challenges can be categorized into three main headings.

6.1 Filtering out Users based on Latency

In the RDG-Image game, latency can potentially affect the collected data in several ways. For example, there can be latency between when a remote user initiates an action in their UI and when the server learns that the action occurred. Pair Me Up includes a latency-measurement protocol that allows for network latency to be monitored and adjusted for (Manuvinakurike and DeVault, 2015). It uses a variant of Network Time Protocol (Mills et al., 2010) to measure the latency. Essentially, ping-pong packets are sent continuously, with timestamps attached, to measure the round trip latency between the client and the server. In (Manuvinakurike and DeVault, 2015), a negative correlation between high mean roundtrip latency and game score was observed. To prevent high latency from affecting this study, we generated 100 such test packets in the **filter users with high latency** step in Figure 7. We then calculated the mean and standard deviation in round trip latency. Users with mean roundtrip latency greater than 250 ms, or with a standard deviation of greater than 45 ms, were filtered out. This helps ensure that latency does not negatively affect the audio channel or gameplay with the agent.

6.2 Dealing with Effects of Variable Latency

Even with the thresholds mentioned in the previous section, transient fluctuations in network latency can sometimes occur, and we found we needed a special mechanism to ensure the integrity of the audio channel. Audio packets are recorded and sent to the PMU server from the client's browser in chunks of approximately 100 ms. Each chunk is sent separately, and is subject to variable transit time due to varying network latency from moment to moment. The order of these packets is thus not guaranteed and they can arrive out of order. For instance, if the audio packets A, B, C are recorded at times t, $t + 100$ms, $t + 200$ms respectively, it is possible for the server to receive them in order A, C, B. If not corrected, this order violation would corrupt the captured audio waveform and potentially degrade ASR and system performance. To overcome this issue, we used an auto-incrementing sequence ID that was appended to each audio packet before it left the user's browser. On the server, we monitor these sequence IDs to make sure that the audio packets either arrive in order or are reordered appropriately by the server.

6.3 Managing Server Load

Even though the agent was designed to handle multiple users at a time, we found in pilot testing that processor and memory usage by the system (agent, webserver, database, ASR) was sometimes too high to support low-latency gameplay by multiple simultaneous users on the available hardware. We therefore decided to limit the agent to one user per server to avoid this issue affecting gameplay, and deployed the system on a commercial cloud-hosting provider using six different servers. Our study could thus support up to 6 simultaneous users. Due to the high attrition rates of participants at various steps in the HIT (Figure 7), sometimes a server was left idle for the maximum HIT completion time of 40 minutes. We did not attempt to build a resource management system to enable more efficient use of our computing resources.

	Web	Lab
Participant Fees	\$1.24	\$15
Staff time per participant	2.5 min	∼35 min
Cost of Server Time	\$0.72/hour/machine	–
Participant Time	1193.1 sec	∼1800 sec

Table 2: Comparison between studies in the lab and web. Estimated numbers are indicated by ∼.

7 Analysis of Crowd-Sourced Study Cost

Table 2 shows several types of measured costs that were incurred in this web-based study (Web column). It also includes, for comparison, an estimate of what the corresponding costs would be for a lab-based human-agent study. The costs in Table 2 for running the study in the lab environment are estimated based on the human-human lab study detailed in (Paetzel et al., 2014).

Participant Fees The web users were compensated an average of \$1.24 (Max=1.56, Min=1.04, SD=0.12) (N=150) per player when interacting with the agent. In the lab, a payment of \$15 was granted for 30 minutes of participation in the human-human study. **Staff time per participant** To manage the HITs on the web required about 2.5 minutes of staff time per participant. In the lab, a staff member needs 30 minutes plus about five more minutes per participant for preparing the lab and the recording equipment.

Cost of Server Time For the 150 successful human-agent participants, the servers in this study were actually used for a total 49.71 hours. The 50 human-human participants required approximately an additional 20 hours of server time. However, due to inefficiencies in our process, during the study, the six servers were kept active for 10 days (1440 server hours). Each server hour costs \$0.72. In the lab, the hardware expenditures for a similar study would be highly dependent on the researcher's environment, but they include the cost of a computer and high-quality audio equipment (about \$800 in our lab).

Participant time The mean total gametime on the web was 275 seconds, but mean participant time was 1193.1 seconds. The additional time was spent by the users on validation steps and answering the questionnaire. In the lab, we estimate that participants would need about 30 minutes for completing the study, including reading and signing the consent form, reading the game rules, playing the game, and answering the questionnaire in the end. In practice, the process takes a little more time in the lab as there is additional time needed for the staff member to greet the participant, manually start the software, adjust the microphone placement, answer any questions, etc.

Over all, it can be seen that this crowd-sourced, web-based approach to human-agent dialogue collection offers potential reductions in several types of costs, including substantial reductions in participant fees and staff time per participant.

8 Limitations

There are several limitations in the way this study was conducted. In the human-human condition, one of the major hurdles is the waiting times involved in creating pairs, which can sometimes be measured in hours (Manuvinakurike and DeVault, 2015). To try to streamline the pairing process, in pilot testing we attempted several methods. We put up a calendar scheduling system where the users could mark their availability, with time slots provided every 30 minutes. Users could avoid waiting to make a pair by selecting a time when another user had stated they were available. However, we found many turkers would select a time slot but then not show up at the specified time. Another technique we tried was a variant of a calendar where the users were paid $0.05 to mark their availability and to then show up at that time. However, again many turkers would not show up at the appointed time. We finally adopted a first-come, first-served method that paired consecutive participants, as was done in (Manuvinakurike and DeVault, 2015). Although this method was relatively slow, as individuals had to wait until a pair could be formed, and had high attrition rates, it was found to work sufficiently well to obtain 25 human-human pairs.

In the human-agent condition, the primary limitation was that there was a large amount of idle system time across our six servers (totaling to about 1370 server hours). This suggests that we had unmet capacity which could have been used to support additional dialogues, or alternatively, we could have used fewer servers to support the same number of users (thus reducing hosting costs). This idle time is related to the high attrition rates (Figure 7) and non-uniform participant presence on AMT during the times when our HITs were active. We aim to tackle these issues by optimizing our HIT and qualification processes in future work.

9 Conclusions and Future Work

In this paper we have reported on a web-based framework that helps address a critical data-collection bottleneck in the design and evaluation of spoken dialogue systems. We demonstrated the viability of our framework through a data collection study in which 200 remote participants engaged in human-human and human-agent dialogue interactions in an image matching game. We discussed several of the technical challenges we encountered and some of the limitations in our current process for collecting dialogue data over the web. In future work, we aim to address the challenge of managing available computing resources better in order to further reduce costs and accelerate data collection.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. IIS-1219253 and by the U.S. Army. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views, position, or policy of the National Science Foundation or the United States Government, and no official endorsement should be inferred. The 8 images of pets used in Figure 1 are excerpted from pictures protected by copyright and released under different licenses by their original authors.[1]

---

[1] Thanks to Joaquim Alves Gaspar for image 1 http://commons.wikimedia.org/wiki/File:Cat_March_2010-1a.jpg and Magnus Colossus for image 3 http://commons.wikimedia.org/wiki/File:Canario_canary_p%C3%A1jaro_bird.jpg, both published under CC BY-SA 3.0. Thanks to Randy Pertiet for image 2 http://www.flickr.com/photos/34652102N04/5428922582/, Brent Moore for image 7 http://commons.wikimedia.org/wiki/File:2006_TN_State_Fair-_Guinea_Pig.jpg and Domenique Godbout for image 8 https://www.flickr.com/photos/dominiquegodbout/5140544743/, all licensed under CC-BY 2.0 and to Opacha for image 4 http://commons.wikimedia.org/wiki/File:Baby–_Yellow_Naped_Amazon_Parrot_Closeup.jpg and TomiTapio for image 6 http://tomitapio.deviantart.com/art/The-bunny-says-nothing-129138755, both licenced under CC-BY 3.0. Thanks to Ilmari Karonen for image 5 https://commons.wikimedia.org/wiki/File:Mouse_white_background.jpg (Public Domain)

References

Ridong Jiang, Rafael E. Banchs, Seokhwan Kim, Kheng Hui Yeo, Arthur Niswar, and Haizhou Li. 2014. Web-based multimodal multi-domain spoken dialogue system. In *Proceedings of 5th International Workshop on Spoken Dialog Systems*.

Walter Lasecki, Ece Kamar, and Dan Bohus. 2013. Conversations in the crowd: Collecting data for task-oriented dialog learning. In *Human Computation Workshop on Scaling Speech and Language Understanding and Dialog through Crowdsourcing*.

Ramesh Manuvinakurike and David DeVault. 2015. Pair Me Up: A Web Framework for Crowd-Sourced Spoken Dialogue Collection. In *International Workshop on Spoken Dialogue Systems (IWSDS)*.

Raveesh Meena, Johan Boye, Gabriel Skantze, and Joakim Gustafson. 2014. Crowdsourcing street-level geographic information using a spoken dialogue system. In *The 15th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*.

D. Mills, J. Martin, J. Burbank, and W. Kasch. 2010. Network time protocol version 4: Protocol and algorithms specification.

Maike Paetzel, David Nicolas Racca, and David DeVault. 2014. A multimodal corpus of rapid dialogue games. In *Language Resources and Evaluation Conference (LREC)*, May.

Maike Paetzel, Ramesh Manuvinakurike, and David DeVault. 2015. "So, which one is it?" The effect of alternative incremental architectures in a high-performance game-playing agent. In *The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDial)*.

Roberto Pieraccini, David Suendermann, Krishna Dayanidhi, and Jackson Liscombe. 2009. Are we there yet? research in commercial spoken dialog systems. In *Text, Speech and Dialogue*, pages 3–13. Springer.

Ondřej Plátek and Filip Jurčíček. 2014. Free on-line speech recogniser based on Kaldi ASR toolkit producing word posterior lattices. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 108–112, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Antoine Raux, Brian Langner, Dan Bohus, Alan Black, and Maxine Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *Interspeech*.

D. Suendermann, J. Liscombe, J. Bloom, G. Li, and R. Pieraccini. 2011. Large-scale experiments on data-driven design of commercial spoken dialog systems. In *Interspeech*.

When Hands Talk to Mouth. Gesture and Speech as Autonomous Communicating Processes

Hannes Rieser
Collaborative Research Center
“Alignment in Communication” (CRC 673)
Bielefeld University, Germany
`hannes.rieser@uni-bielefeld.de`

Abstract

The implementation of speech-gesture interfaces is one of the vital problems in formal research on multi-modal discourse. This paper provides empirical evidence that, due to asynchronous occurrences of gesture and speech, speech-gesture interfaces cannot be expressed in purely static structural terms resting on a speech-gesture map. As an alternative, a methodology is suggested which models gesture and speech as independent communicating processes generating together multi-modal content. It is based on a dynamic process algebra, the π-calculus. To meet the descriptive needs of speech-gesture interface construction, the π-calculus is extended to a hybrid λ-π-calculus devised to handle higher order information.

1 Introduction

Recently, there has been a growing interest to investigate the coordination of gesture and speech in multi-modal discourse, originally initiated by scholars like McNeill (1992) and Kendon (2004). I’ll take up this topic in my paper as well, providing a new approach. The leading idea, motivated by corpus studies in sec. 5, is that speech and gesture work as independent processes, abstract agents which communicate and together produce a multi-modal content, e.g., if a winding gesture modifies the word “street”. It will be shown in due course that this also necessitates a move from classical algorithmic modelling, be it λ-calculus, Montague Grammar (MG) or some brand of Dynamic Semantics, to process modelling using dynamic calculi such as the π-calculus. I will substantiate this idea in the following way: Sec. 2 starts with some assumptions embodying the idea to take gesture and speech as independent processes. This is further motivated in sec. 3. Sec. 4 presents McNeill’s influential observations on speech-gesture coordination. In sec. 5 examples of static speech-gesture interfaces are provided resting mainly on McNeill’s ideas. Sec. 6 elaborates on three case studies showing asynchrony of gesture and speech yielding counter examples to static speech gesture interfaces. Sec. 7 comes with intuitive process analyses for the asynchrony cases involving parallel processes and process-interaction. In sec. 8, I introduce a process algebra, namely π-calculus, show how to extend it to a hybrid of λ-π-calculus, and use the resulting machinery for the description of gesture-speech interaction. I close with some indications for future research in sec. 9.

2 Assumptions

I assume that speech and gesture have meaning, say along the lines of a Peircean semiotics. As a consequence, I take it that speech meaning and gesture meaning can be represented and computed independently but that there is some coordination between them. This is obvious, e.g., from demonstrations accompanying the use of indexical expressions: demonstrations have to be coordinated with the production of the indexical (Lücking et al., 2015). The *locus* of speech-gesture coordination is informally called “interface” here, following common practice in software engineering to compute information of different type from different sources. In the interface, speech information and gesture information are stored and processed.

3 Idea of the paper

As is obvious from the remarks on the interface, the interface between speech meaning and gesture meaning has to be expressed in a formal way. The question is then which formalism to use. The options are data structures suitable to interface information. So it is no surprise that Mc-

Neill (1992) used frame-structures (reconstructed in Röpke (2011)). A more recent concept resorts to AVMs in an HPSG-representation (Lücking, 2013). In general, I assume that the modelling of gesture must be based on rigid (rated) annotation, annotation playing for gestures the same role as syntax representation plays for linguistic utterances. Speech acts and gesticulations are widely different types of structure, there is no natural mapping from one to the other comparable to a syntax-semantics-map. As will be shown in the case studies below (sec. 5), natural speech gesture interface data resist modelling in conventional structural terms (such as trees, AVMs or pure FOL-representations). As a consequence, so I argue, one must look for different conceptualizations. The main problems for a natural mapping from gesture to speech are that the gesture often does not exactly overlap its fitting speech counterpart: It comes too early, too late or extends over too much language material. So, there is no semantic synchrony. A machinery which seems to be able to capture this dynamics at least partially are process algebras such as the π-calculus (Parrow (2001) and Sangiorgi and Walker (2001)), the Calculus of Communicating Systems (CCS, Milner (1999)) or Communicating Sequential Processes (CSP, Hoare (1985)). Using one of these will move one from an object and proposition metaphysics to one based on processes.

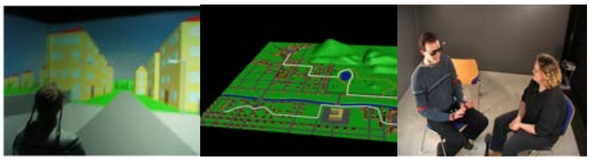


Figure 1: The virtual town, the route traversed, Route-giver and Follower sitting in the cave.

The dynamics of the speech-gesture relation will be shown using the speech and gesture alignment corpus, SaGA (Lücking et al., 2013). It contains 25 route description dialogues from three camera perspectives. The dialogue participants are a Route-giver and a Follower, the Route-giver explaining his/her route through a virtual town to a Follower. Lücking et al. gathered video and audio data, body movement tracking data, and eye-tracking data. Approx. 7500 gestures have been identified, 6000 of them annotated and rated. Due to the experimental setting, they have to deal with the genre of multi-modal task-oriented dialogue with many specific dialogue structures, such as clarification sequences, repetitions and tests.

4 Mc Neill on Speech-gesture Coordination

McNeill (1992) using the so-called Tweety-data was the first scholar to provide generalizations on speech-gesture coordination which are widely used for interface construction, although, as I will show below, his approach is in the end too normative and prone to falsification. Here I provide McNeill's semantic synchrony rule (McNeill (1992), p. 27) and his definition of stroke (McNeill (1992), p. 83) for further use.

Semantic synchrony rule: Semantic synchrony [of gesture and speech, author] means that the two channels, speech and gesture, present the same meanings at the same time. The rule can be stated as follows: "if gesture and speech co-occur they must cover the same 'idea unit'" [i.e., content, author].

Stroke: Stroke [. . .] is the peak effort in the gesture. It is in this phase that the meaning of the gesture is expressed. The stroke is synchronized with the linguistic segments that are co-expressive with it.

From McNeill's definitions of semantic synchrony and stroke it follows that the set-up of a speech-gesture interface is provided by (the content of) the gesture's stroke and the meaning of the synchronised linguistic material. These two have to interact. For example, an iconic gesture indicating a square can interact with the semantics of, say, "envelope", indicating the envelope's shape.

5 Static speech-gesture interfaces: frames and HPSG-matrices

McNeill (1992) was interested in specifying the generation of speech-gesture ensembles as shown in fig. 2. The important issue here is that a filled frame is used to store the information necessary for generating speech and (optionally) an accompanying gesture. We get the information needed packed into one static data structure. A more recent variant of a static data structure is provided by Lücking (2013) who uses an HPSG-grid to model speech-gesture interfaces (fig. 3).

The relation under discussion on this grid is a two-dimensional "round", "round2", which gets

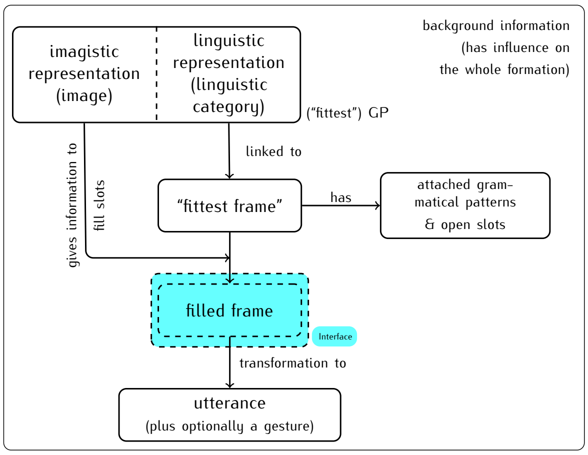

imagistic representation (image)
linguistic representation (linguistic category)
background information (has influence on the whole formation)
("fittest") GP
linked to
gives information to
fill slots
"fittest frame"
has
attached grammatical patterns & open slots
filled frame
Interface
transformation to
utterance (plus optionally a gesture)

Figure 2: McNeill's speech-gesture generation frame as reconstructed in Röpke (2011).

its semantics from a TRAJectory in the G(esture)-DaughTeR as is evident from the unification $\boxed{7}$. As in the frame case, we have a static structure. The trajectory's semantics can enter into exactly one position of the RESTRiction of "round2". The set-up of the speech-gesture ensemble is quite powerful due to unification but we cannot go into details here, check especially $\boxed{3}$ and $\boxed{5}$. The content of fig. 2 and fig. 3, respectively, might well serve as a kind of *explicans* to the McNeill quotes above. Mainly for didactic reasons I have chosen Lücking's approach as a prototypical one here but I think that the same arguments apply tot he HPSG-based speech gesture interfaces of Alahverdzhieva and Lascarides (2010) who also use structure-based technologies. Furthermore, work in the SDRT-tradition focusing on the explication of gesture meaning is based on similar interface conceptions (Lascarides and Stone (2006) and (2009)) and faces similar falsifying instances. Hopefully, the difference to the process-based proposals made in this paper will become clear from the following case studies (sec. 6) and the process analyses in sec. 7. How the findings presented here carry over to incremental theories of information in the manner of, e.g. Hough et al. (2015), still remains to be investigated, however, the suspicion is that they do carry over. Turning to the point of view of hypothesis falsification the question arises whether we find gesture-speech occurrences where speech and gesture belong together intuitively but do not obey McNeill's synchrony rule. Below I present the essentials of three case studies showing exactly such falsifying instances. They also serve as falsifying instances for static speech-gesture interfaces.

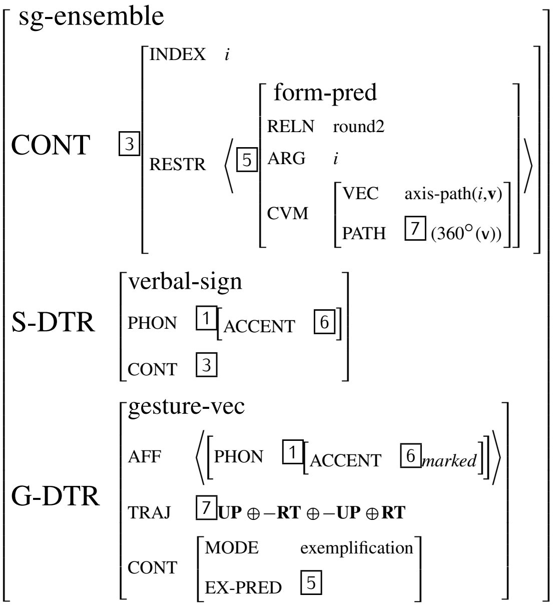
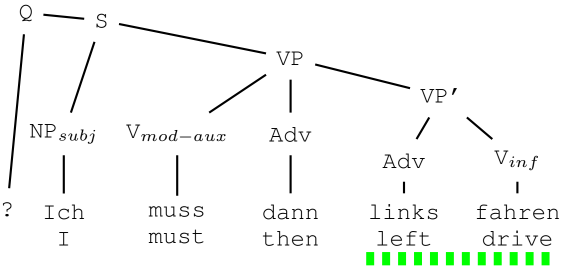


Figure 3: HPSG-grid to model speech-gesture interfaces from Lücking (2013), p. 249.

6 Three case studies: Asynchrony of gesture and speech (based on Hahn and Rieser (2012))

In the following, intuitive notions like "channel", "communicate", "interaction", "interfacing", "process" or "sending" are used. They are given a proper algorithmic reconstruction in sec. 8.

6.1 Case I: Indexing is held too long



(a) Syntax of Follower's clarification request. The stroke is marked with a green dashed line.

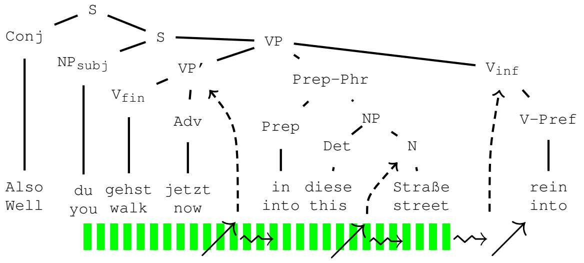
$$\begin{bmatrix} \text{Hand\_Shape} & \textit{L} \\ \text{Back\_Of\_Hand\_Direction} & \textit{BTL} \\ \text{Palm\_Direction} & \textit{PTB} \\ [\ldots] & \end{bmatrix}$$

(b) AVM of gesture annotation.

Figure 4: Datum of Case I.

Fig. 4 shows a Route-giver and Follower exchange, the syntax of the clarification request *?I must then left drive* and the annotation of the Follower's gesture, a demonstration to the left. The green marks indicate the gesture stroke overlap with *left* and *drive*. According to McNeill the stroke should only overlap with *left*. Hence, the Follower's indexing is held too long. At first sight an explanation could be given which is in agreement with McNeill, namely, if we interface the gesture stroke with the VP'. However, doing that we would lose bottom-up compositionality, because the terminal "left" is not related to the stroke.

6.2 Case II: Object gesture must wait to compose



(a) Syntax of Follower's clarification request. The stroke is marked with a green dashed line.

Hand_Shape	*B-spread*
Back_Of_Hand_Direction	*BAB/BTL>BTL>BAB* >BAB/BTL
Palm_Direction	*PTB/PTL>PTB>PTL* >PTB/PTL
Wrist_Movement_Direction	*MF/ML/MU*

(b) AVM of gesture annotation.

Figure 5: Datum of Case II. Stroke of gesture overlapping several constructions, *inter alia*, the subject NP.

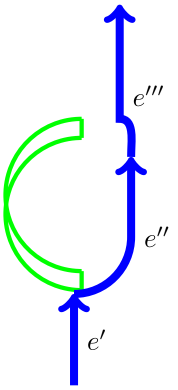
Fig. 5 shows a clarification request of a Follower. Again I provide the gesture annotation and the syntax structure, here of the 2nd utterance. The green marks represent the stroke of the winding gesture. Observe that the winding information is not contained in the utterance, so we have additional information in the gesture. Although there are several options for speech-gesture interfaces, the preferred *locus* of integration is "street", yielding winding street in a multi-modal way, whereas it could not easily be combined with "walk now" or "into". Since the stroke starts overlapping with "you", we have again a counter-example to McNeill's synchrony rule.

6.3 Case III: Multi-parallelism and anaphora

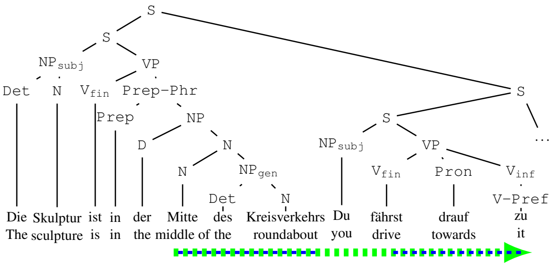
```
Es ist aber    auch ein Kreisverkehr.
It is  however also a   roundabout.
Die Skulptur  ist in der Mitte  des
The sculpture is  in the middle of the
Kreisverkehrs. Du  fährst drauf   zu,
roundabout.    You drive  towards it,
rechts herum     und dann  ja    und
right  around it and then, well, and
dann quasi geradeaus      sozusagen
then quasi straight-ahead so to speak
die  abbiegen.
this one leaving.
```

(a) Route-giver's directive.

LH	
Hand_Shape	*C-loose*
Back_O_H_Direction	*BAB*
Palm_Direction	*PTR*
RH	
Hand_Shape	*H-loose*
Back_O_H_Direction	*BAB*
Palm_Direction	*PDN*
Wrist_Mov_Direction	*MF*



(b) LH- and RH-annotation and Trajectories e', e'', e''' representing the trajectories of "drive towards", "right around it", and "leaving it", respectively.



(c) Syntax tree and stroke overlaps as dashed lines, left hand green, right hand blue.

Figure 6: Datum of Case III.

Fig. 6 represents the Route-giver's utterance, the annotation of the left hand, the right hand and the trajectories e', e", and e"'. The "natural interface" in these cases is not marked by a gesture-stroke speech overlap. A more elaborate description of the right-hand and the left-hand activities and their relation to speech will be given in the next section.

6.4 Generalisation

Let us generalise from the findings in the case studies: Given that we have at least two information channels, an alternative to static speech-gesture interfaces emerges: We model the respective information on two channels and how they communicate. Still, after having dealt with the issues due to the falsification, we must be aware of the fact that at the ultimate speech gesture contact points, i.e., if we have successfully singled out the appropriate interface, we will encounter problems as those indicated by McNeill, Lücking, Rieser (2013) and others, namely, how to represent the alignment of speech meaning and gesture meaning. This shows that these researchers discovered something important but used an idealised case prototypes of which can also be found in the data.

7 Process analyses for the asynchrony cases

This section contains intuitive analyses of the case studies. They are informally expressed and serve as a sort of precondition for the discussion about communicating processes in sec. 8.

7.1 Taking up case I: Indexing is held too long

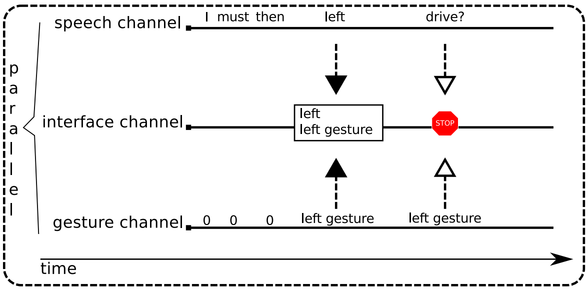


Figure 7: Three parallel channels: RH, LH, interface and speech channel.

Fig. 7 depicts three channels operating in a parallel way. On the speech channel we have the utterance “I must then left drive”?. On the gesture channel there are first empty events indicated by 0. The interface channel encodes the interaction between the information on the speech channel and the information on the gesture channel. It also indicates where the interfacing can occur (boxed area) and where it can’t. Accordingly, the semantics of the word “left” and the left gesture interact in some time interval but there is no interaction afterwards, indicated by the red stop sign. So, the idea is to restrict the activity of the gesture, more precisely, that of its semantic representation, to the word “left”. Observe that the gesture meaning itself is in no ways annihilated; it remains on the gesture channel.

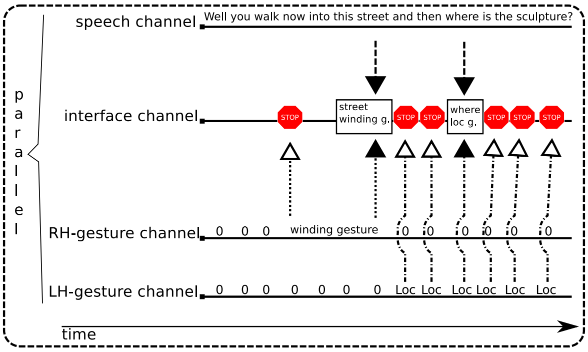


Figure 8: Four parallel channels: RH, LH, interface and speech channel.

7.2 Taking up case II: Object gesture must wait to compose

As fig. 8 shows, we have a RH- and a LH-gesture channel, both interacting with the speech channel which transports “Well you walk now into this street and then, where is the sculpture”?. Here the RH’s gesture comes too early at “now” which cannot combine with the winding. Since it continues to send *via* the extended gesture stroke, it can finally cooperate with “street” yielding in the end the multi-modal semantics ⟦winding street⟧. The LH gesture starts to communicate when speech contributes “and” and “then” on the speech channel. However, to become effective, it has to wait until “where” turns up, then providing indexical information for it in the gesture space (producing Quinean deferred reference). After that the speech’s cooperation potential is used up and the LH-gesture is barred off from contribution to the interface channels.

7.3 Taking up case III: Multi-parallelism and anaphora

Fig. 9 shows that we have various interfaces active. LH and RH first communicate to produce a cylindrical shape and the shape information then tries to get access to the speech level. The example also shows the use of a linguistic anaphora “the sculpture” and of an anaphora at the gesture level. More on that further down. In more detail,

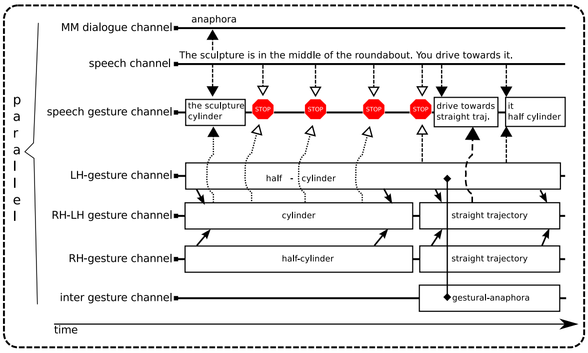


Figure 9: RH- and LH-gesture channel communicate forming the "composite" RH-LH gesture channel. This channel communicates with the speech channel on the speech-gesture channel. In addition, there is a multi-modal dialogue channel on which linguistic anaphora ride and an inter-gestural channel for gestural anaphora.

the LH-gesture and the RH-gesture each form a half-cylinder and together shape a cylinder. The cylinder-information can communicate with the speech information "the sculpture" on the speech-gesture channel. Then the cylinder information is stopped from interacting with speech. Afterwards the LH and the RH part company. The RH shapes a straight trajectory whereas the LH still signs the half-cylinder, forming a "gestural anaphora" for the whole cylinder. However, LH and RH start to cooperate on the RH-LH-interface channel. Together they indicate a situation involving a cylindrical object (LH) and a path around it (RH) which can contribute to the meaning of "drive towards" which clearly involves a target. The LH continues to send information, contributing after at least one stop indexical meaning to the anaphora "it".

7.4 Evaluation of data and development of the formal mechanisms needed to describe flexible speech gesture-interaction

If we look at the speech gesture interaction, we find that actions like "stop, I do not want information" (indicated by the red stop sign), processes and process interactions seem to be the most basic entities. We encountered different types of processes, empty ones, speech-gesture, gesture-gesture. Processes run in parallel as our timelines indicate. They hook up to each other *via* interfacing. They emit or receive information. In the case studies gesture is as a rule the emitting source and speech the receiver. Receiving can imply that processes are changed by information, remember the multi-modally specified winding street. However, information can also be neglected or blocked. Processes can be recursive, this can be seen, when a process tries to communicate several times (thus generating a daughter process of the same type) but is barred from the interface. Interactions among channels come in sequences. Clearly, we need an algorithm which can capture at least some of that.

8 From λ- to π-calculus. The step to process algebra

Before we deal with processes, we enter a familiar field: the λ-calculus. Formal work in NL semantics often relies on applied λ-calculus. It has logical constants, constants for individuals and relations, operators for all styles of variables plus the λ-operator. It often works with a generalised quantifier representation and rules of $\alpha\beta\eta$-conversion (see Curry et al. (1974), p. 92). It has inspired semantic work from Church and Curry to Montague and beyond. In contrast, the π-calculus' basic entities are names, represented by lower case letters. They are used by processes/channels ("channel" now being a technical term) for interactions. Interactions have to be formally indicated: Capabilities for actions are provided by so-called *prefixes* π:

$$\pi \; := \; \overline{x}y \mid x(z) \mid \tau \mid [x = y]\pi$$

Then we have processes P and summations M:

$$P \; := \; M \mid P \mid P' \mid \nu z P \mid !P$$

$$M \; := \; 0 \mid \pi.P \mid M \; + \; M'$$

Among the prefixes we have an output prefix $\overline{x}y$ and an input prefix $x(z)$. $\tau.P$ evolves invisibly to P. There is a match prefix, $[x = y]$, in $[x = y]\pi.P$. In a sum $P + P'$ either P or P' can be executed but not both. "$P|P'$" is called "composition". In such a composition, P and P' can be executed independently, in parallel or interact *via* shared names, yielding output-input devices. Shared names are already indicated in π above. $\nu z P$ states that the scope of name z is restricted to process P, in traditional parlance, z is treated much like a bound variable. $!P$ denotes infinite iteration, defined as $P|!P$, i.e., a process composed with an iteration of processes.

Given our intuitive analyses of asynchrony cases in sections 6 and 7, what do we get from the π-calculus? First of all, a technical nomenclature for our intuitive distinctions like process etc. (see the list in sec. 7) and then algorithmic means for modelling them. In more detail: We have parallel channel modelling *via* composition "|". As already indicated, there are output-input devices *via* the prefixes $\overline{x}y$ (outputting y) and $x(z)$ (receiving a name *via* x and substituting it for z in the subsequent process). We have types of binding, tests and arbitrarily deep recursion due to replication. In addition, type systems for channel names can be given, a device which we will exploited below.

8.1 Typing and a hybrid λ-π-machinery

So, the advantages of π seem to be fairly clear. But hold on! Essentially, we would like to have the expressive power of the higher order λ-calculus in the interfaces, gesture-gesture and gesture-speech alike, as we have seen them in the asynchrony studies. The reason is that the information seems to be higher order. My suggestion is to achieve that through

(a) using λ-operator and π-operator based definitions for λ-calculus names a, b, etc. resulting in mixed λ-π-expressions

(b) using typed λ-calculus constants as π-calculus names for channels, λ-calculus constants are given the status of π-calculus names

(c) letting channels have a MG type such as e, $\langle e,t\rangle$, $\langle\langle e,t\rangle,\langle e,t\rangle\rangle$, etc. in order to match them with type-fitting names.

I call the π-calculus extended with (a), (b), (c), hybrid λ-π-calculus.

We are now equipped to handle interfacing speech-gesture processes and turn to an illustrative example. We do the relatively simple case I from section 7, indexing is held too long (cf. fig. 4). For didactic reasons, I use the English word by word translation "I must then left drive?" and reconstruct the crucial "left drive" in λ-π-calculus terms. A further simplification is added: Because of operators "must" and "then" I use a more tractable version of the English translation, namely "Must then I left drive?" I provide first the speech representation, then the speech interface representation and the definitions of names and types for the λ-π-calculus representation (cf. table 1). Holding indexing is modelled by "$!P$". Parallel channels of speech and gesture are modelled by "|".

Speech representation	Speech interface representation	Types and definitions for π-calculus names
$I := \lambda P.P(I)$	same	$\overline{x'}, x', \overline{z}, u, w$: MG-type $< e, t >$
drive $:= \lambda x.\text{drive}'(x)$	same	$\overline{x}, x, y$: MG-type $<< e, t >, < e, t >>$
left $:= \lambda f \lambda z.\text{left}'(f)(z)$	$\lambda z.\text{left}'(w)(z) \wedge y(w)(z)$	$a := \lambda z.\text{left}'(w)(z) \wedge y(w)(z)$
must $:= \lambda p.\Box p$	same	
then $:= \lambda p \lambda q.\text{then}'(p, q)$	same	
Gesture representation	Gesture interface representation	
$!\overline{x}\text{left}'.0$	same	
Speech-gesture interface	gesture representation	speech representation
$x(y).x'(w).\ a.0 \mid$	$!\overline{x}\text{left}'.0 \mid$	$\overline{x'}\text{drive}'.0$

Table 1: Speech representation, gesture representation and their λ-π-interface.

As in the figures, green colouring indicates gesture information. Replication definition for $!\overline{x}\text{left}'.0$ yields:

$$x(y).x'(w).\ a.0 \mid \overline{x}\text{left}'.0 \mid\, !\overline{x}\text{left}'.0 \mid \overline{x'}\text{drive}'.0 \tag{1}$$

We substitute π-names in the processes by their λ-π-definitions and get:

$$x(y).x'(w).\lambda z.\text{left}'(w)(z)\ \wedge\ y(w)(z).0 \mid \overline{x}\text{left}'.0 \mid\, !\overline{x}\text{left}'.0 \mid \overline{x'}\text{drive}'.0 \tag{2}$$

$\overline{x}$ outputs left$'$ to x: y is instantiated with left$'$:

$$x'(w).\lambda z.\text{left}'(w)(z)\ \wedge\ \text{left}'(w)(z) \mid 0 \mid\, !\overline{x}\text{left}'.0 \mid \overline{x'}\text{drive}'.0 \tag{3}$$

$\overline{x'}$ outputs drive$'$ to x': w is instantiated with drive$'$:

$$\lambda z.\text{left}'(\text{drive}')(z)\ \wedge\ \text{left}'(\text{drive}')(z) \mid 0 \mid\, !\overline{x}\text{left}'.0 \mid 0 \tag{4}$$

We get the property "$\lambda z.\text{left}'(\text{drive}')(z)\ \wedge\ \text{left}'(\text{drive}')(z)$" which after normalization becomes "$\lambda z.\text{left}'(\text{drive}')(z)$".

Observe that "compositionally used up" information results in 0-events. For some reasons (perhaps to facilitate coherence, to add emphasis or to maintain the focus) the gesture is kept on its channel. This is what "$\overline{x}\mathrm{left}'.0$" expresses. Again, this is additional information for separating the gesture channel and the speech one. So what we get in the end is an algorithmic rendering of the intuitive representation in fig. 7.

So, the gesture does not contribute new content to the speech content. But, while the word "left" evaporates, the indexing on the gesture channel is still visible as we have it in the datum and the diagram fig. 10. It can still be SEEN when the next word "drive" is already HEARD, leading to a division of labour among channels.

8.2 A note on generalisability

Finally, a word on generalisability of the λ-π-calculus account might be in order: We need multi-channel renderings in various multi-modal contexts anyway, take, e.g., tone-group information not strictly co-extensive with syntax trees, the information contained in eye-tracking data or in body postures. So, multi-channel representations seem to be an imperative research venue to follow.

9 Conclusion and further research

As shown in the case studies, in the SaGA data speech portions and gesture strokes do not perfectly synchronize. We have seen that grammar-based speech gesture interfaces cannot deal with gestures produced too early, lagging behind or intruding" into "alien" speech material by, e.g., crossing propositional boundaries, expressing contradictory content etc. As a way out we propose to consider speech and gesture as autonomous concurrent processes communicating with each other *via* an interface. This can be achieved by exploiting the facilities of the suggested λ-π-calculus to model higher order properties of concurrent speech-gesture events and gesture-gesture events.

As the λ-π-hybrid shows, we have lost some of the pleasant simplicity of the pure π-calculus. It might also not be evident at first sight what the inductive definition of the λ-π-hybrid would look like, due to the mixture of λ-names and π-variables in a single expression. Certainly, some problems remain, but having concentrated in this paper on the defence of using process algebras for the description of multi-modal discourse, we defer these matters to a more theoretical paper on the λ-π-hybrid.

Acknowledgements

Thanks go to Florian Hahn and Insa Lawler for work on gesture during the last years. Support by the German Research Foundation in the CRC 573, "Alignment in Communication", Bielefeld University, is also gratefully acknowledged.

References

Katya Alahverdzhieva and Alex Lascarides. 2010. Analysing language and co-verbal gesture in constraint-based grammars. In Stefan Müller, editor, *Proceedings of the 17th International Conference on Head-Driven Phase Structure Grammar (HPSG)*, pages 5–25, Paris.

Haskell B. Curry, Robert Feys, and William Craig. 1974. *Combinatory Logic*, volume 1 of *Combinatory Logic*. North-Holland Publishing Company, 3 edition.

Florian Hahn and Hannes Rieser. 2012. Non-compositional gestures. In *International Workshop on Formal and Computational Approaches to Multimodal Communication held under the auspices of ESSLLI 2012*, Opole.

Charles A. R. Hoare. 1985. *Communicating sequential processes*. Prentice-Hall International series in computer science. Prentice-Hall, Englewood Cliffs, NJ, 6. print edition.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 206–216, London, UK. Association for Computational Linguistics.

Adam Kendon. 2004. *Gesture – Visible Action as Utterance*. Cambridge University Press, Cambridge, NY.

Alex Lascarides and Matthew Stone. 2006. Formal semantics of iconic gesture. In David Schlangen and Raquel Fernández, editors, *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, Brandial'06, pages 64–71, Potsdam. Universitätsverlag Potsdam.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: the bielefeld speech and gesture alignment corpus (saga) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Andy Lücking, Thies Pfeiffer, and Hannes Rieser. 2015. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79.

Andy Lücking. 2013. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. De Gruyter Mouton, Germany.

David McNeill. 1992. *Hand and Mind. What Gestures Reveal About Thought*. The University of Chicago Press.

Robin Milner. 1999. *Communicating and Mobile Systems: The π Calculus*. Cambridge University Press, Cambridge.

J. Parrow. 2001. An introduction to the π-calculus. In A. Ponse J.A. Bergstra and S.A. Smolka, editors, *Handbook of Process Algebra.*, pages 479–545. Elsevier.

Hannes Rieser. 2013. Speech-gesture interfaces. an overview. In *Proceedings of the 35th Annual Conference of the German Linguistic Society (DGfS)*, pages 282–283.

Insa Röpke. 2011. Watching the growth point grow. In *Proceedings of the Second Conference on Gesture and Speech in Interaction (GESPIN) 2011.*

D. Sangiorgi and D. Walker. 2001. *The π-calculus. A Theory of Mobile Processes.* Cambridge University Press, Cambridge.

Interpreting English Pitch Contours in Context

Julian J. Schlöder
Institute for Logic, Language & Computation
University of Amsterdam
`julian.schloeder@gmail.com`

Alex Lascarides
School of Informatics
University of Edinburgh
`alex@inf.ed.ac.uk`

Abstract

This paper presents a model of how pitch contours influence the illocutionary and perlocutionary effects of utterances in conversation. Our account is grounded in several insights from the prior literature. Our distinctive contribution is to replace earlier informal claims about the implicatures arising from intonation with logical derivations: we validate inferences in the SDRT framework that resolve the partial meaning we associate with a pitch contour to different specific interpretations in different contexts.

1 Introduction

In this paper, we give a formal semantics of pitch contour in spoken dialogue, implemented in SDRT (Asher and Lascarides, 2003). Our main claim is that the pitch contour of an utterance conveys cognitive attitudes, and that taking these attitudes into account perturbs calculable implicatures. The following examples, adapted from Steedman (2000), are cases in point:[1]

(1) A: You're a millionaire.
a. B: I'm a MILLIONAIRE.
H* LL%
b. B: I'm a MILLIONAIRE?
H* LH%
c. B: I'm a MILLIONAIRE?
L* LH%

(2) A: Are you rich?
a. B: I'm a MILLIONAIRE?
H* LH%

The utterance in (1a) is an assertion with the 'high focus, final fall' contour (H* LL%). Conventionally, this commits B to the proposition *'B is a millionaire'* and thereby establishes agreement between A and B. The same utterance with a final rise (LH%) in (1b) is a question (*'Am I?'*) that does not make any such commitment. The low pitch accent (L*) in (1c) additionally reveals that B is somewhat surprised or doubtful about A's assertion (*'Am I? Really?'*). While (2a) has exactly the same form as (1b), in the context of (2) B *does* make a commitment to being a millionaire, but displays uncertainty on whether this answers A's question. These examples show that the intonation of an utterance can influence both illocutionary and perlocutionary inferences: in (1b) an indicative mood utterance is a clarification request, and (1c) expresses the failure of belief transfer after an assertion. They also show that such inferences are highly context-sensitive.

Our formal account makes specific, computable predictions on what attitudes are displayed by pitch, and when a particular pitch contour is licensed. It achieves this by leaving the compositional semantics of pitch deliberately underspecified, with contextual information and inference supporting a specific and complete interpretation in context. We believe that our model is novel in its formal precision, with previous work resorting to semi-formal paraphrases of how intonation gives rise to implicatures.

Steedman (2014) formalises pitch contours in terms of their effect on common ground, and claims that the effects outlined above are derivable from general principles of truth maintenance. However, he does not give a formal account of these derivations. Our model proposes a compositional semantics for individual pitch accents in terms of public commitment; our semantic postulates are inspired by Steedman's, but we *formally derive* their specific contribution in context. However, we abstract away from grammatical parsing and assume that a grammar is in place which connects to our semantics of pitch. This means that we do not take the lexical placing of the focus accent into account and assume that the foreground proposition of an utterance is computed elsewhere.

In the next section, we expand our informal discussion to further examples. We give a brief introduction to the formal framework of SDRT in section 3, including some amendments to SDRT's cognitive logic. We present our formal theory in

---

[1] To describe our examples, we use the ToBI annotation scheme throughout (Silverman et al., 1992).

section 4 and show that it corresponds to the analyses surveyed in section 2. In section 5 we conclude and give pointers towards further work.

2 Informal Discussion

The meaning of pitch contours in English has received substantial attention in the literature. We briefly review some prevalent discussions and use them to motivate our formal model. The data we present is not new: it is derived from a number of earlier analyses that comprehensively survey the phenomena we are interested in. Where we constructed an example for the sake of exposition, we verified that our reading is in accordance with earlier accounts.

2.1 Final Rise

We follow the discussion of Schlöder (2015). A typical interpretation of the final rise in English is that it signals 'insufficiency' in various senses. Hobbs (1990) and Bolinger (1982) characterised the meaning of a final rise as signalling 'incompleteness.' Šafářová (2005) has characterised this incompleteness as displaying, *i.a.*, an 'uncertain attitude.' Westera (2013) has further specified such uncertainty by relating it to the Gricean maxims: the speaker displays uncertainty regarding the truthfulness (Quality), specificity (Quantity) or appropriateness (Relation) of their utterance. Given these observations, we characterise the final rise as marking an utterance as incomplete, but consider incompleteness itself to be an underspecified notion, *i.e.*, incompleteness can be resolved in different ways. The following two possible continuations, adapted from Hirschberg and Ward (1995), exemplify this variation in resolution:

(3) a. A: Where are you from?
b. B: I'm from SKOKIE.
H* LH%
c. B: That's in Illinois.
c.' A: Okay, good.

The final rise in (3b) is interpreted to signal incompleteness. Specifically, B is uncertain if the answer in (3b) pragmatically resolves A's question. Both follow-ups in (3c) and (3c') resolve this incompleteness in different ways: in (3bc), B is supplying additional information himself, signalling that *'Skokie'* is not the full answer, but that *'Skokie, Illinois'* is. In (3bc'), B is waiting for A to comment on whether she considers (3b) to be sufficient. Thus in (3bc'), B's utterance is taken to have question force (in some sense). However, it cannot be glossed as *'am I [from Skokie]?'* Rather, it is *'I'm from Skokie—does that answer your question?'* So here, the incompleteness is resolved in the dialogue as posing a question that needs to be answered by A. Similarly, the final rise in (1b) is interpreted as a clarification request. However, the illocutionary force of (1b) *can* be paraphrased as *'Am I [a millionaire]?'* So a proposition p presented in indicative mood with final rise can have the illocutionary effect of the polar question $?p$ (*'Is it the case that* p*?'*) in some but not all contexts.

Hence, our basic idea to describe incompleteness goes as follows: the final rise always *projects* a follow-up, *i.e.*, it demands some kind of resolution, but we leave open what kind of speech act is being projected and what precisely its contribution to the discourse is (*cf.* Malamud and Stephenson (2011)). We then infer the specific illocutionary force of the final rise utterance and of its follow-up via a logic of defeasible inference that draws on the compositional and lexical semantics of the utterances involved. Where appropriate, we default to clarification questions, *i.e.*, our logical axioms validate a defeasible inference that the preferred follow-up to (1b) is a confirmation. But we leave room to resolve the projection in another way in different contexts, *e.g.*, by elaboration in (3bc) or acceptance in (3bc').

2.2 Pitch Accents

We are primarily interested in the meaning of a low pitch accent (L*). A typical interpretation of the low pitch in English is that it indicates that belief transfer has failed in some way: Hobbs (1990) takes L* to indicate that the foreground proposition is either known or false, and Steedman (2000; 2014) takes it to mean that grounding fails. We are also interested in the low pitch with high border accents (H+L* and L*+H). Consider the following possible responses in (4):

(4) A: France has a king!
a. B: France is a MONARCHY.
L* LL%
'It is not, this is obvious.'
b. B: France is a MONARCHY.
H+L* LL%
'It is not, this is obvious.'
c. B: # France is a MONARCHY.
L*+H LH%

The utterances in (4a) and (4b) are intonated to openly display sarcasm, and are hence disagreeing with A's assertion. However, using the so-called contradiction contour (*cf.* Liberman and Sag (1974)) as in (4c) is incoherent. In dialogues where—abstracting away from pitch—the semantics of B's response contradicts A's, the felicity and meaning change:[2]

(5) A: France has a king!
a. B: France is a REPUBLIC.
L* LL%
'It is obviously a republic.'
b. B: # France is a REPUBLIC.
H+L* LH%
c. B: France is a REPUBLIC.
L*+H LH%
'You are wrong, it is a republic.'
d. B: # France is a REPUBLIC.
H* LH%
e. B: # France is a REPUBLIC.
H* LL%

Unsurprisingly, the contradiction contour in (5c) is licensed now. The L* pitch that earlier gave rise to a sarcastic reversal of meaning now yields a rather condescending correction in (5a), but (5b) is incoherent. Perhaps unexpectedly, a high pitch is also incoherent here. Clearly, the question in (5d) cannot be a clarification request, but the case for the incoherence of (5e) is more opaque: speaker A asserts something, and B asserts the opposite. *Prima facie* this is alright, but with these pitch contours it seems as if A and B are talking past each other.

We conclude that while the low pitch generally indicates a problem with belief transfer, its meaning depends on its surrounding border accents: the simple low pitch (L*) signals disagreement irrespective of the utterance's actual propositional meaning, whereas the complex pitches (H+L* and L*+H) change meaning (and felicity).

Hobbs (1990) defines the effect of H* as marking the foreground proposition as 'new.' Since 'not new' to him means grounded or false, 'new' should denote ungrounded *and* true. Since, in dialogue, truth *as far as the participants are concerned* takes precedence, we paraphrase Hobbs' 'new' instead as 'uncontroversial.' Steedman's (2014) gloss of the H* LL% contour is *'I succeed in making common ground that p.'* We are reluctant to accept that a speaker can make a public commitment for *another* speaker (*pace* Gunlogson (2003)). Hence, one cannot *individually* succeed in making something common ground, and we can only consider Steedman's gloss to mean that the speaker *assumes* that belief transfer will succeed, *i.e.*, that he has no reason to believe otherwise. We again can paraphrase this as 'p is uncontroversial.' This link from H* to 'uncontroversial' explains the incoherence of (5d) and (5e): after A commits to France having a king, the proposition *'France is a republic'* cannot be considered uncontroversial.

Example (6), taken from Ladd (1980), shows that intonation has more subtle effects when we are interested in more nuanced differences than assent *vs.* contradiction.

(6) A: Harry's the biggest liar in town.
a. B: The biggest FOOL maybe.
H* LL%
b. B: The biggest FOOL maybe.
L*+H LH%

The contour (6a) has B agreeing with A's assertion by *elaborating* it, *i.e.*, B is publicly committed to Harry being both the biggest liar *and* maybe the biggest fool. The contour in (6b) does not commit B to Harry being the biggest liar, but it is not an outright denial either. Instead, B is committed to Harry being the biggest fool, and that this makes *'Harry is the biggest liar'* less believable. Hence, the utterance in (6b) has a different illocutionary force than (6a), and attaches to the discourse as *counterevidence*.

While the H* LH% contour in (7a) again displays uncertainty, the examples in (7b,c) reveal something new about the low pitch: B conveys something about how his beliefs have changed *in the aftermath* of A making her previous utterance:

(7) A: Did you read the first chapter?
a. B: I read the THIRD chapter?
H* LH%
'Does that suffice?'
b. B: I read the THIRD chapter?
L*+H LH%
'Wasn't I supposed to read the third?'
c. B: I read the whole DISSERTATION.
L*+H LL%
'And you should know that I did.'

In (7b), B believed it to be common ground that he is to read the third chapter, and in (7c) he believed that A knew that he had read the whole dissertation. To make sense of these perlocutions, a listener needs to draw inferences of the form *'before A said u, B must have believed p.'* Such reasoning is also useful to describe the meaning of contours signalling surprise. In the next section, we describe how we formalise such *hindsights* in SDRT's cognitive modelling logic.

---

[2] The glosses/interpretations of the utterances in (4) and (5) are our own. We additionally verified that they are consistent with Steedman's (2014) explanations of the grounding-related effects of pitch contours.

3 Framework

Our theory of pitch contours is implemented in Segmented Discourse Representation Theory (Asher and Lascarides, 2003; Lascarides and Asher, 2009). Our rationale goes as follows. SDRT models back-and-forth information flow between three interconnecting languages and associated logics: the *language of information content*, the *glue logic*, and the *cognitive modelling logic*. By manipulating this information flow, we gain fine-grained control over different aspects of an utterance's interpretation, allowing us to model perturbations of the standard interpretations. Each of the logics in SDRT is designed for a specific task, and we briefly describe each of them in turn.

The *language of information content* is used to express the logical form of a discourse, capturing its pragmatically resolved, specific interpretation. The dynamic semantics of this language models the truth conditions of the public commitments that speakers make through their utterances. The language of information content includes *rhetorical relations* (*e.g.*, $Explanation$ or $Elaboration$) that connect the representations of individual discourse units.

A logical form in SDRT is an SDRS: a set of labels Π, where each label stands for a discourse segment, and a mapping $\mathcal{F}$ from each label in Π to a formula representing that segment's information content (we will sometimes write $\mathcal{F}(\pi)$ as K_π). Since these formulae include rhetorical relations among labels, $\mathcal{F}$ imposes an ordering on Π: π_1 *immediately outscopes* π_2 if $\mathcal{F}(\pi_2)$ features $R(\pi_1, \pi)$ or $R(\pi, \pi_2)$ as one of its conjuncts. We write $\pi \succeq \pi'$ for the transitive closure of this relation. A well-formed SDRS imposes the constraint that this partial order is rooted, *i.e.*, there is a single discourse segment consisting of rhetorically connected subsegments.

In dialogue, participants make *public commitments* to SDRSs. Specifically, the logical form of a dialogue turn is a set of SDRSs, one for each dialogue participant. When a speaker utters a unit π, he commits to the *rhetorical relation* that connects π to the prior context. In effect, this makes speakers publicly committed to the illocutionary contribution of their moves. The logical form of a dialogue is the logical forms of its turns. For example, the logical form of (8) is as follows:

(8) A: Max fell.
B: John pushed him.

Turn	A's SDRS	B's SDRS
1	$\pi_1 : \mathit{fall}(e, m)$	$\emptyset$
2	$\pi_1 : \mathit{fall}(e, m)$	$\pi : \mathit{Explanation}(\pi_1, \pi_2)$ $\pi_2 : \mathit{push}(e', j, m)$

The dynamic semantics of $Explanation(\pi_1, \pi_2)$ entails the contents of π_1 and π_2 in dynamic conjunction, and that the latter answers the question *'why is* π_1 *true?'* Being publicly committed to $Explanation(\pi_1, \pi_2)$ thus makes B publicly committed to the content of π_1. This means that A and B agree that Max fell—they share a public commitment to it—even though this is an implicature of B's contribution and not linguistically explicit.

There are several parts of the logical form of (8) that go beyond the compositional and lexical semantics of its individual units: the pronoun *'him'* is resolved to m, and the illocutionary contribution of B's utterance is to provide an *Explanation* to A's. These inferences are about the *construction* of logical forms (as opposed to their *truth*). As input, they take *underspecified logical forms* (ULFs), which are in turn computed from an utterance's linguistic surface form. This construction is modelled in the *glue logic* which we discuss next.

The glue logic validates defeasible inferences from partial descriptions of logical forms (*i.e.*, ULFs) to fully specified discourses (*i.e.*, SDRSs like those in 8). SDRSs capture the pragmatically preferred and complete interpretation of the discourse. These inferences are facilitated by axioms of the following form:

$$(\lambda :?(\alpha, \beta) \wedge \mathit{Info}(\alpha, \beta)) > \lambda : R(\alpha, \beta).$$

The $>$ denotes a default conditional and we use Greek letters to label discourse segments. So, informally, the above formula says: 'if α and β are rhetorically connected to form a part of the extended discourse segment λ, and their ULFs satisfy *Info*, then normally, their rhetorical connection is R.' Such default axioms are justified by word meaning, world knowledge and cognitive states.

The default conditional $>$ yields a nonmonotonic proof theory $\mid\!\sim_G$. To ensure that the glue logic remains decidable, it reasons about ULFs (*i.e.*, the (partial) *form* of a logical form), but has only limited access to what those logical forms *mean* in the logic of information content. Keeping the glue logic decidable accounts for how people by and large agree on *what* was said, if not on whether it is true.

The glue logic also has access to information in the *cognitive modelling logic*. This logic includes a number of modal operators: KD45 modal operators for beliefs (B_S for a speaker S); K45 operators for public commitment (P_S); and special modal operators for intentions (I_S).[3] Also, for each action term δ, there are two modal operators $[\delta]$ ('after δ') and $[\delta]^{-1}$ ('before δ'); see Asher and Lascarides (2008) for a discussion of this logic. The only action term we will be concerned with is the act of uttering something, $\delta = s_S(\pi)$ for an utterance label π and its speaker S. For our purposes, these operators are sufficiently specified by postulating the following axioms:

Glue to Cognitive Logic (GL to CL).
Let $\pi_1 \dots \pi_n$ be elementary discourse units spoken by $S_1 \dots S_n$, and Γ_n be the context after π_n (*i.e.*, their ULFs plus facts and axioms). Let $\vdash_G$, $\mathrel{|\!\sim}_G$ be the monotonic and nonmonotonic proof theories of the glue logic. Let $\vdash_C$ and $\mathrel{|\!\sim}_C$ be the ones for the cognitive modelling logic.
If $\Gamma_n \vdash_G \varphi$, then

$$\Gamma_n \vdash_C [s_{S_1}(\pi_1)] \dots [s_{S_n}(\pi_n)] P_{S_n} \varphi.$$

If $\Gamma_n \mathrel{|\!\sim}_G \varphi$, then

$$\Gamma_n \mathrel{|\!\sim}_C [s_{S_1}(\pi_1)] \dots [s_{S_n}(\pi_n)] P_{S_n} \varphi.$$

Persistence.
If $\Gamma \mathrel{|\!\sim}_C P_A \varphi$ and $A \neq S$, then $\Gamma \mathrel{|\!\sim}_C [s_S(\pi)] P_A \varphi$. A person's public commitments are unaffected by another speaker's utterance.

Hindsight.
If $\Gamma_n \mathrel{|\!\sim}_C [s_{S_1}(\pi_1)] \dots [s_{S_n}(\pi_n)][s_{S_i}(\pi_i)]^{-1} B_S \varphi$, then $\Gamma_n \mathrel{|\!\sim}_C [s_{S_1}(\pi_1)] \dots [s_{S_{i-1}}(\pi_{i-1})] B_S \varphi$. 'Before'-operators cancel up to a corresponding 'after'-operator.

Conservativity.
$([s_S(\pi)] B_{S'} \varphi) \rightarrow (B_{S'} \varphi \vee B_{S'}((P_S K_\pi) > \varphi))$. Beliefs after an utterance are either carried over from before, or are inferred from that utterance.

Reduction.
$(B_{S'}[s_S(\pi)]\varphi) > ([s_S(\pi)] B_{S'} \varphi)$, and $(B_{S'}[s_S(\pi)]^{-1}\varphi) > ([s_S(\pi)]^{-1} B_{S'} \varphi)$. Beliefs usually transfer to hindsight and foresight judgements, *i.e.*, if a speaker believes that after/before the act π, the proposition φ holds, they have that belief in foresight/hindsight.

The axioms *GL to CL* and *Persistence* together ensure that glue logic inferences about the illocutionary act that a speaker performs matches their (current) public commitments in the cognitive logic; so if A has asserted that p then in the cognitive logic A is publicly committed to p. Conversely, defeasible inferences made in the glue logic can also be *blocked* by *facts* from the cognitive modelling logic, *e.g.*, if $\Gamma \vdash_C P_S \neg\varphi$, then the glue logic cannot defeasibly infer a discourse relation in S's SDRS that would entail φ.

Note that the context Γ_n in *Hindsight* does not change. The axiom models inferences that interlocutors can make about previous cognitive states *from their current knowledge* Γ_n, which *extends* their prior knowledge Γ_{i-1}. In particular, it is possible that the axiom applies in Γ_n, but that $\Gamma_{i-1} \not\mathrel{|\!\sim}_C [s_{S_1}(\pi)] \dots [s_{S_{i-1}}(\pi_{i-1})] B_S \varphi$. Also note that the hindsight-inferences formalised by the *Hindsight* and *Reduction* axioms are scoped by a belief modality. Since defaults support belief revision (*i.e.*, it is possible that $\Gamma \mathrel{|\!\sim} B_S \varphi$ while $\Gamma \wedge \psi \mathrel{|\!\sim} B_S \neg\varphi$), the above axioms support revision *in hindsight*. We go more in-depth on these phenomena in the next section.

4 Formal Model of Pitch Contours

We now give a precise, formal account of the effects we discussed in section 2. We first give a brief account of cooperative principles in SDRT and how they are used to compute the perlocutionary effects of utterances. This initial presentation will discuss the standard (unperturbed) inferences. We then present our semantics for pitch contours, and afterwards show how we derive their pragmatic effects.

4.1 The Standard Reasoning

Our main concern are the *perlocutionary* contributions of pitch contours, which we model in SDRT's cognitive modelling logic. In SDRT, such effects (like belief transfer) are specified by stipulating axioms affecting the cognitive models of the speakers (Asher and Lascarides, 2003; Asher and Lascarides, 2013). The following axioms give a Gricean account of cooperativity:

Sincerity (a). $P_S \varphi > B_S \varphi$.
Sincerity (b). $B_S \neg\varphi > \neg I_S P_S \varphi$.
Cooperativity. $P_S I_S \varphi > I_H \varphi$.
Intention Transfer. $P_S \varphi > P_S I_S P_H \varphi$.

Sincerity states that public commitments are usually truthful regarding the interlocutor's beliefs, *Cooperativity* that publicly announced intentions are usually adopted by their addressee, and *Intention Transfer* that a public commitment is usually intended to be grounded, *i.e.*, to become a shared public commitment. In SDRT, both interlocutors

---

[3] Glossing over the details, we write $I_A B_B \varphi$ if A wants B to believe that φ, and $I_A P_B \varphi$ if A wants B to commit to φ.

maintain their own private model of the cognitive modelling logic, *i.e.*, their individual representation of the public commitments, beliefs and intentions of everybody involved in the conversation. We assume that everyone agrees on the above axioms, and that this fact is mutually known.

As an example, suppose that a speaker S asserts p to a hearer H. By *GL to CL*, S and H infer $P_S p$ in the cognitive model. Then, H can infer that S actually believes that p by *Sincerity*. Further, both can infer by *Intention Transfer* that S wants H to make the same commitment, *i.e.*, $P_S I_S P_H p$. By *Cooperativity* the speaker S can infer that $I_H P_H p$ and so expects an agreement move (establishing H's commitment to p) next.

4.2 Final Rise

Based on our discussion in section 2, we take the final rise to have an influence on: (i) the *structure* of the dialogue by demanding a follow-up (incompleteness); (ii) the illocutionary force of an utterance (*e.g.*, an inferred question force); and (iii) the inferred attitudes of the speaker (uncertainty). We refine the model of Schlöder (2015). The following mapping formalises incompleteness:[4]

Semantics of the Final Rise.

$$\pi(LH\%) \mapsto \pi' = ? \wedge \pi'' = ? \wedge R = ? \wedge R(\pi', \pi'') \wedge \pi' \succeq \pi.$$

That is, the final rise semantics enforces that there is a yet unknown follow-up response standing in some relation to the final rise unit π. We leave open what discourse relation is projected, and we allow it to attach to a wider discourse segment as long as it includes π as a part. This is required to model cases where it is the *discourse relation itself* that is uncertain. For example, in (3bc'), where A accepts the *Question-Answer-Pair* (QAP) relation itself (*i.e.*, that 3b answers her question 3a), and not just the contents of (3b). That is, A's move is $\mathit{Accept}(\pi, c)$, where $\pi : \mathit{QAP}(a, b)$. In (3bc), however, the projected relation is $\mathit{Elaboration}(b, c)$, directly attaching to the final rise utterance (3b).

In addition, we stipulate a glue logic axiom that, where truth-conditionally appropriate, defeasibly infers from a final rise that a question is being asked. The following rule serves to interpret an indicative mood utterance with content p as the polar question $?p$ (as in example 1a):[5]

Clarification from Final Rise.

$$\big(\beta : LH\% \wedge \lambda : ?(\alpha, \beta) \wedge \Box(K_\alpha \rightarrow \mathit{prop}(K_\beta))\big) > \lambda : \mathit{CR}(\alpha, \beta).$$[6]

In this axiom, $\pi : LH\%$ means that the label π includes the final rise semantics. So this axiom stipulates that if an utterance has a final rise, and its core propositional content is entailed by that of its attachment point, then normally, it is a clarifying polar question.[7] The entailment $K_\alpha \rightarrow \mathit{prop}(K_\beta)$ is required to explain the incoherence of (9b):

(9) A: You are rich.
a. B: I'm rich? *'Am I?'*
b. B: # I'm a millionaire?

(10) A: You are a millionaire.
a. B: I'm rich? *'Am I?'*
b. B: I'm a millionaire? *'Am I?'*

Both answers in (10) are licensed because, conventionally, *'millionaire'* implies *'rich,'* hence the question in (10a) is reasonable. Conversely, *'rich'* does not necessarily imply *'millionaire,'* so B's utterance in (9b) does not support an interpretation as a clarification request. Lastly, we model 'uncertainty' in SDRT's cognitive modelling logic. Here, the functions $S(\pi)$ and $H(\pi)$ map a label to its speaker and hearer, respectively.

Cognitive Contribution of the Final Rise.

$$\pi : LH\% \wedge \lambda : R(\alpha, \pi) \wedge \neg\pi : ?\mathit{prop}(K_\pi) > P_{S(\pi)} \neg B_{S(\pi)} I_{H(\pi)} P_{H(\pi)} R(\alpha, \pi).$$

This stipulates that if the utterance with the final rise directly attaches to an antecedent, but is not a question,[8] then the speaker publicly displays uncertainty about whether the hearer is willing to commit to that relation. This is in particular true if the relation is *Correction* (as in, *e.g.*, 5c), but also applies to uncertain answers as in (3b).

As discussed in section 4.1, the combined application of *Cooperativity* and *Intention Transfer* would normally yield $I_{H(\pi)} P_{H(\pi)} R(\alpha, \pi)$, *i.e.*, the hearer will establish a shared commitment on the discourse relation R in the next turn. The cognitive contribution of the final rise conveys that the speaker S was unable to make that inference—for whatever reason. We take this to be the underspecified uncertainty that a final rise communicates.

---

[4] *cf.* Pierrehumbert and Hirschberg (1990): 'to interpret an utterance with particular attention to subsequent utterances.'

[5] Note that if an utterance is in interrogative mood, then the axioms of Asher and Lascarides (2003) will already support an inference that the utterance has the force of a question.

[6] CR $\simeq$ Clarification Request. CRs have question semantics, *i.e.*, $\pi : ?K_\pi$, and are *sincere* (not rhetorical): $\neg P_S K_\alpha \wedge \neg P_S \neg K_\alpha$. *CR* has the dynamic semantics of *elaborating questions* (Asher and Lascarides, 2003, p. 468).

[7] It is necessary to map K_β to its propositional content, as once question force is inferred, K_β is a question.

[8] On questions, the final rise is part of the default contour and cannot be taken to convey uncertainty.

4.3 Pitch Accents

We only discuss the cognitive functions of nuclear pitch accents, abstracting away from pre-nuclear pitches and lexical position. We are furthermore only concerned with pitch accents on indicative utterances (including those interpreted as questions), but not with interrogatives. We stipulate the following cognitive contributions (we simplify notation by setting $S = S(\pi)$, $H = H(\pi)$):

Cognitive Contributions of Nuclear Accents.

- $\pi(H^*) \mapsto P_S(\neg B_S B_H \neg K_\pi)$.
  'I don't think what I'm saying is controversial.'
- $\pi(H{+}L^*) \mapsto P_S(\neg I_S P_S K_\pi)$.
  'I'm not committing to what I just said.'
- $\pi(L^*) \mapsto \lambda : ?(\alpha, \pi) \to \Big(P_S\big(\neg I_S P_S K_\alpha\big) \wedge P_S\big([s_H(\alpha)]^{-1} B_S \neg I_H P_H K_\alpha\big)\Big)$.
  'I didn't think you'd want to commit to what you just said, and I'm unwilling to.'
- $\pi(L^*{+}H) \mapsto \lambda : ?(\alpha, \pi) \to \Big(P_S\big(B_S B_H \neg K_\pi\big) \wedge P_S\big([s_H(\alpha)]^{-1} \neg B_S B_H \neg K_\pi\big)\Big)$.
  'I didn't think what I'm saying is controversial, but now I do.'

Note that the postulate for H* states that the speaker S assumes that belief transfer, as formalised by the successive application of *Intention Transfer* and *Cooperativity*, will succeed. To be precise, if S's cognitive model would include $B_S B_H \neg K_\pi$, then S would infer by *Sincerity (b)* that $B_S \neg I_H P_H \varphi$, and would hence believe that *Cooperativity* would not apply, *i.e.*, S would not expect an agreement move next. Intonating H* is therefore the default pitch insofar that S explicitly communicates that she sees no reason why the standard grounding process should *not* obtain, yielding the implicature *'I expect you to agree.'* Such an expectation is unwarranted if the utterance is a correction move; in section 4.4 we show how this explains the incoherence of (5e).

Conversely, the first conjunct of the L*+H contribution has the speaker conveying the opposite: she assumes that her utterance's content *is* controversial. Accordingly, the L*+H contour features prominently in utterances that put two propositions in contrast, *e.g.*, in denials. We give a formalisation of example (6b) in section 4.4.

In the first clause of the L* contribution, however, a speaker is explicitly announcing that the *Cooperativity* axiom has failed on *their* side of the model, and that belief transfer on H's earlier statement (labelled α) has failed. The cognitive contribution of the H+L* pitch has the same form, but relates to the current utterance (labelled π): the speaker is indicating that she does not intend that her own utterance's contents be grounded. [9] If the propositional content of π is the same as that of α, the result is a sarcastic rejection (as in 4b). Usually, such a rejection is taken to mean that the speaker actually believes the opposite. Hence we include the following negation-strengthening axiom:

Sarcasm. $P_S \neg I_S P_S \varphi > P_S \neg \varphi$.

This reads as follows: if someone makes the explicit public commitment to *not* make a particular commitment, they are usually taken to commit to the opposite. This accounts for the actual reversal of meaning in a sarcastic utterance, instead of a mere refusal to ground. In the next section, we show how this axiom separates the sarcastic rejection (4a) from the doubtful question (1c).

What is left to discuss are the second clauses of the L* and L*+H contributions, respectively. These clauses convey something about earlier beliefs, allowing for hindsight inferences. By uttering something, a speaker incurs a public commitment. The second clause of the L* contribution conveys that the next speaker did not expect this commitment. The second clause of the L*+H contribution relates to an utterance's content being thought uncontroversial, but that belief having now changed—thereby allowing inferences on the speaker's beliefs *before* the utterance in hindsight.

4.4 Applications

We now verify derivations of the effects of pitch contours for four of our earlier examples.

Ex. (5e) A: France has a king!
B: # France is a REPUBLIC.
H* LL%

From A's utterance we can infer:

$\Gamma \vdash [s_A(\alpha)][s_B(\pi)]\Box(K_\alpha \to \neg K_\pi)$ (fact).

$\Gamma \vdash [s_A(\alpha)] P_A K_\alpha$ (GL to CL).

$\Gamma \mathrel{|\!\sim} [s_A(\alpha)][s_B(\pi)] P_A K_\alpha$ (Persistence), hence

$\Gamma \mathrel{|\!\sim} [s_A(\alpha)][s_B(\pi)] P_A \neg K_\pi$.

$\Gamma \mathrel{|\!\sim} [s_A(\alpha)][s_B(\pi)] B_A \neg K_\pi$ (Sincerity)

$\Gamma \mathrel{|\!\sim} B_B [s_A(\alpha)][s_B(\pi)] B_A \neg K_\pi$
(axioms are mutually believed).

---

[9] *cf.* Steedman (2014) *'I fail to make it common ground.'*

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)][s_B(\pi)]B_B B_A\neg K_\pi$ (Reduction).

From B's utterance, including its H*, we get:

$\Gamma\vdash[s_A(\alpha)][s_B(\pi)]P_B\neg B_B B_A\neg K_\pi$

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)][s_B(\pi)]B_B\neg B_B B_A\neg K_\pi$ (Sincerity).

$\Gamma\mathrel{|\!\sim}[s_H(\alpha)][s_B(\pi)]\neg B_B B_A\neg K_\pi$ (B is KD45[10]).

Hence we infer that one of A or B are insincere. Since it is indeterminable to an overhearer *who* is insincere, *i.e.*, which application of Sincerity is blocked, the dialogue appears incoherent.

Ex. (6b) A: Harry's the biggest liar in town.
B: The biggest $\underset{\text{L*+H}}{\text{FOOL}}$ maybe$\underset{\text{LH\%}}{.}$

The intended reading is that B is putting his utterance in contrast to A's utterance. We start with the second conjunct of the L*+H semantics:

$\Gamma\vdash[s_A(\alpha)][s_B(\pi)]P_B\big([s_A\alpha]^{-1}\neg B_B B_A\neg K_\pi\big)$.

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)][s_B(\pi)]B_B\big([s_A\alpha]^{-1}\neg B_B B_A\neg K_\pi\big)$ (Sincerity).

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)][s_B(\pi)][s_A\alpha]^{-1}B_B\neg B_B B_A\neg K_\pi$ (Reduction).

$\Gamma\mathrel{|\!\sim}B_B\neg B_B B_A\neg K_\pi$ (Hindsight).

$\Gamma\mathrel{|\!\sim}\neg B_B B_A\neg K_\pi$ (B is KD45)[11] $(*)$.

Now, the first conjunct of the model for L*+H:

$\Gamma\vdash[s_A(\alpha)][s_B(\pi)]P_B\big(B_B B_A\neg K_\pi\big)$.

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)][s_B(\pi)]B_B\big(B_B B_A\neg K_\pi\big)$ (Sincerity).

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)][s_B(\pi)]B_B\big(B_A\neg K_\pi\big)$ (B is KD45).

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)]\big(B_B B_A\neg K_\pi \vee B_B(P_B K_\pi > B_A\neg K_\pi)\big)$ (Conservativity).

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)]B_B B_A\neg K_\pi$ ($\vee$-elimination).[12]

$\Gamma\mathrel{|\!\sim}B_B\big(P_A K_\alpha > B_A\neg K_\pi\big)$ (Conservativity + $*$).

$\rightsquigarrow$ *'That you told me he is a liar tells me that you don't think he is a fool.'*

Ex. (4a) A: France has a king!
B: France is a $\underset{\text{L*}}{\text{MONARCHY}}\underset{\text{LL\%}}{.}$

By the first conjunct of the model for L*:

$\Gamma\vdash[s_A(\alpha)][s_B(\pi)]P_B\neg I_B P_B K_\alpha$.

$\Gamma\mathrel{|\!\sim}[s_A(\alpha)][s_B(\pi)]P_B\neg K_\alpha$ (Sarcasm).

$\rightsquigarrow$ *'You are wrong.'*

This inference channels back into the glue logic, which now validates the discourse relation $\mathit{Correction}(\alpha,\pi)$, entailing $\neg K_\alpha$, in B's SDRS.

Also, similar to the derivation of $(*)$, applying Sincerity, Hindsight and Reduction to the second conjunct yields:

$\Gamma\vdash[s_A(\alpha)][s_B(\pi)]P_B[s_A(\alpha)]^{-1}B_B\neg I_A P_A K_\alpha$.

$\Gamma\mathrel{|\!\sim}B_B\neg I_A P_A K_\alpha$.

$\rightsquigarrow$ *'I thought you wouldn't say that.'*

In sum, B is communicating to A in (4a) that he is correcting A and that he did not expect that he would have to do so.

Ex. (1c) A: You're a millionaire.
B: I'm a $\underset{\text{L*}}{\text{MILLIONAIRE}}\underset{\text{LH\%}}{?}$

Here, B's utterance has question force, but it is read with a bias towards the negative answer. First, the axiom *Clarification from Final Rise* renders B's utterance as a CR. Hence, the cognitive contribution of the final rise does not apply. Now, consider the second conjunct of the L* contribution and apply, as before, Reduction and Hindsight:

$\Gamma\mathrel{|\!\sim}B_B\neg I_A P_A K_\alpha$

$\rightsquigarrow$ *'I thought you wouldn't say that'* ($\approx$ surprise).

Then, by the first conjunct of the model for L*:

$\Gamma\vdash[s_A(\alpha)][s_B(\pi)]P_B\neg I_B P_B K_\alpha$.

$\rightsquigarrow$ *'I'm unwilling to agree with what you just said.'*

In contrast to (4a), Sarcasm cannot be applied here, because it is blocked by the dynamic semantics for clarification requests (CRs must be sincere questions). Hence, A is not communicating that B is *wrong*, but just that B is not ready to agree.

5 Conclusion

We have presented a unified, formal account of the perlocutionary effects of pitch contours in colloquial English as discussed in the literature. The novel contribution of our model is the formal derivability of these effects. Our stipulations of cognitive effects are independently motivated and in line with previous analyses of these effects. By connecting them with the logics of SDRT, we obtain concrete derivations of implicatures communicated by pitch. In future work, we plan to extend this analysis to interrogatives and imperatives. Further, we have ignored the focus effects of the lexical placement of pitch accents here. To integrate these effects into our account, we plan to extend SDRT's glue and cognitive logics to reasoning with the contents of sub-clausal units.

---

[10] KD45 models introspection, *i.e.*, $B_B\varphi \rightarrow B_B B_B\varphi$. Hence, if $B_B\neg B_B\varphi$, then $B_B\varphi$ must fail.

[11] The analoguous derivation for examples (7b) and (7c) accounts for the *'I thought you knew'* implicature.

[12] By Int. Transfer+Cooperativity, $P_B K_\pi\mathrel{|\!\sim} I_A P_A K_\pi$, and by Sincerity (b), $B_A\neg K_\pi\mathrel{|\!\sim}\neg I_A P_A K_\pi$, hence the second disjunct normally does not apply.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher and Alex Lascarides. 2008. Commitments, beliefs and intentions in dialogue. In *12th Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL)*.

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6(2):1–62.

Dwight Bolinger. 1982. Intonation and its parts. *Language*, pages 505–533.

Christine Gunlogson. 2003. *True to form: Rising and falling declaratives as questions in English*. Routledge.

Julia Hirschberg and Gregory Ward. 1995. The interpretation of the high-rise question contour in english. *Journal of Pragmatics*, 24(4):407–412.

Jerry Hobbs. 1990. The pierrehumbert-hirschberg theory of intonational meaning made simple: Comments on pierrehumbert and hirschberg. In J. Morgan P. R. Cohen and M. E. Pollack, editors, *Intentions in Communication*, pages 313–323. MIT Press.

D Robert Ladd. 1980. *The structure of intonational meaning: Evidence from English*. Indiana University Press.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26(2).

Mark Liberman and Ivan Sag. 1974. Prosodic form and discourse function. In *10th Regional Meeting of the Chicago Linguistics Society*, pages 416–427.

Sophia A Malamud and Tamina Stephenson. 2011. Three ways to avoid commitments: declarative force modifiers in the conversational scoreboard. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, pages 74–83. Los Angeles.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In J. Morgan P. R. Cohen and M. E. Pollack, editors, *Intentions in communication*, pages 271–311. MIT Press.

Julian J Schlöder. 2015. A formal semantics of the final rise. In *Student Session of the European Summer School in Logic, Language and Information 2015*.

Kim EA Silverman, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti Price, Janet B Pierrehumbert, and Julia Hirschberg. 1992. ToBI: a standard for labeling english prosody. In *2nd International Conference on Spoken Language Processing*.

Mark Steedman. 2000. *The syntactic process*. MIT Press.

Mark Steedman. 2014. The surface-compositional semantics of english intonation. *Language*, pages 2–57.

Marie Šafářová. 2005. The semantics of rising intonation and interrogatives. In E. Maier, C. Bary, and J. Huitink, editors, *Proceedings of SuB9*, pages 355–369.

Matthijs Westera. 2013. Attention, im violating a maxim! A unifying account of the final rise. In *17th Workshop on the Semantics and Pragmatics of Dialogue (DialDam)*.

“Hand me the yellow stapler” or “Hand me the yellow dress”: Colour overspecification depends on object category

Sammie Tarenskeen
s.tarenskeen@gmail.com

Mirjam Broersma
m.broersma@let.ru.nl

Bart Geurts
brtgrts@gmail.com

Radboud University
Nijmegen, The Netherlands

Abstract

Two production experiments were conducted to investigate how colour overspecification varies with the object category the referent falls into. We found a positive correlation between how important colour is for objects and how likely speakers are to produce colour overspecification when referring to those objects. We also found that speakers tend to produce colour overspecification when referring to geometrical figures, even though colour is considered of low importance for this category. Following Arts et al. (2011a) and Koolen et al. (2011), we assume that speakers tend to include colour because it is often a highly salient attribute of objects. We argue that on the one hand, colour importance increases colour salience, accounting for the correlation between colour importance and colour overspecification, and on the other hand, the paucity of other attributes of simple figures increases colour salience, accounting for the high proportions of colour overspecification for this category. We claim that variation in colour overspecification across object categories is due to the general cooperative strategy of including salient attributes, which are helpful in referent identification.

1 Introduction

When speakers use definite descriptions to refer to objects and individuals, they have to select information about the referent. How speakers do this is currently a major question in research on reference (van Deemter et al., 2012). We present a series of studies that provide insight into this question. We investigate how characteristics of object categories increase the degree to which colour is salient for objects in those categories, and hence the likelihood that speakers select colour when referring to them. For example, are speakers more likely to select colour when referring to a dress, the colour of which is important and therefore presumably salient, than when referring to a stapler? And do they select colour more often when referring to simple figures such as circles and squares, which have no other attributes than colour and shape to attract the attention, than to more complex, real life objects? We focus on colour *overspecification*, which occurs when colour is included even though a unique description of the referent does not require mention of colour in the particular visual context, e.g., ‘the red dress’ in a context where the referent is the only dress. Although reference has many functions in dialogue, we limit ourselves to the function of referent identification.

Early theories assumed that speakers tend to select those attributes with highest discriminatory power, that is, attributes which distinguish between the referent and most of the other objects in the context, thereby avoiding attributes that are not necessary for the addressee to identify the referent (Ford and Olson, 1975). Experiments, in contrast, have shown repeatedly that speakers do not tend to produce such minimal descriptions of their referents, but often include unnecessary attributes, resulting in overspecification. Moreover, the discriminatory power of attributes does not seem to be a significant factor in the selection process (Gatt et al., 2013; Viethen et al., 2014); instead, speakers have preferences for certain attributes. In particular, speakers seem to prefer mentioning colour, sometimes selecting it even when it has no discriminatory power at all, that is, when all objects in the context share the referent’s colour (Koolen et al., 2013). Colour is included more often without need than other attributes, like size (Belke, 2006), material (Sedivy, 2005), and location (Arts et al., 2011b). That is, overspecification is most often *colour* overspecification.

Why is colour preferred so strongly? The common view is that colour is a salient property of objects (Arts et al., 2011a; Koolen et al., 2011). We can think of several reasons why this might be so. Colour is used to identify objects and to distinguish between objects: it is a basic cue in interpreting our visual image (Treisman and Gelade, 1980). It is also an absolute attribute (Pechmann, 1989; Belke and Meyer, 2002): to determine the colour of an object, it need not be compared to other objects, in contrast to determining whether it is big or small. Eyetracking data suggest that speakers often start to articulate colour adjectives even before looking at other objects, whereas they only include size after detecting a size difference between the referent and another object of the same type (Brown-Schmidt and Konopka, 2011). We suggest, then, that colour is visually highly accessible. It is also linguistically accessible: many languages have a fine-grained colour lexicon, which enables speakers to easily label virtually all colours they can perceive and to use unique labels for a wide variety of colours (Berlin and Kay, 1969). Colour is probably special in being both visually and linguistically more accessible than most if not all other attributes.

It is sometimes argued (Engelhardt et al., 2006) that overspecification is in conflict with Grice's theory of pragmatics (Grice, 1975). After all, the second maxim of quantity should prevent us from producing an utterance that is more informative than is required. This line of argument does not seem to do justice to the Gricean framework. Grice's point is not that we obey to a set of (stipulative) rules; it is that communication is a form of *cooperative* behaviour. If including information into a referring expression is not necessary but nevertheless helpful in the identification of the referent, it is an act of cooperativeness to do so (Arts et al., 2011a). It is a good idea for a cooperative speaker to mention an attribute that is salient to her: such an attribute is likely to be salient to the addressee too, and therefore helpful in referent identification. Salient attributes are not necessarily required for the ultimate purposes of the discourse, but including them does improve the efficiency of the comprehension process. Indeed, it has been found that overspecification can result in shorter referent identification times than minimal descriptions (Mangold and Pobel, 1988; Arts et al., 2011a).

This is not to say that speakers *always* produce the referring expressions that are optimal for comprehension. Language production is constrained by the way our cognitive system is organised, and producing an expression that is optimal for the addressee can therefore be inefficient for the speaker. The smoothness of the exchange may thus improve if an expression is produced that is more efficient from the speaker's point of view and suboptimal but nevertheless understandable from the addressee's point of view. There is evidence that unnecessary information can hinder the comprehension process (Altmann and Steedman, 1988; Engelhardt et al., 2011; Davies and Katsos, 2013) and it is an interesting empirical question in what situations hearers detect overspecification, and what happens when they do. It seems a reasonable assumption, however, that colour is such a helpful cue in referent identification that, in general, addressees do not tend to detect the redundancy of colour overspecification and are usually not hindered by it.

In this paper, we focus on characteristics of objects that contribute to the colour salience of those objects. Of course, characteristics of the visual context contribute to colour salience of objects, too. The colour of a blue object, for instance, is less salient if all other objects in the context are blue. Hence, we would expect speakers to be less likely to produce colour overspecification in such contexts than when the referent is surrounded by objects in different colours, which is indeed what has been found (Koolen et al., 2013). This finding is readily explained in the Gricean framework: it is likely that an addressee detects the redundancy of a colour adjective and is hindered by it when all objects surrounding a blue referent are also blue. Another way in which colour salience is affected by the visual context is when the colour of an object is atypical for this type of object: the colour of a purple crocodile is arguably more salient than the colour of a green crocodile, and we would expect the probability that colour overspecification is produced to increase correspondingly. This too has been confirmed by experimental data (Westerbeek et al., 2014). Again, an explanation in the Gricean framework is easily provided: a hearer who is not told about the colour of a purple crocodile will initially look for a green individual to no avail, and when he has identified the referent, the question why the speaker did not mention such a salient feature may confuse him even further. Producing colour overspecification is then more cooperative than avoiding it.

If it is true that speakers have a general tendency to include salient attributes – which is generally compatible with cooperative behaviour – patterns in attribute selection may occur that are *not* readily explained in terms of cooperativeness. As has been suggested before, for example, colour is intuitively not equally important for all object categories (Rubio-Fernández, 2011): most people will presumably consider colour more important for fashion items than for construction tools. If higher colour importance increases colour salience, we would expect that speakers are more likely to produce colour adjectives and colour overspecification when referring to a fashionable bag than to an electric drill, all else being equal. Yet, 'red' is probably not more helpful in identifying the referent when it is a bag than when it is a drill. Selecting colour when referring to a bag but not when referring to a drill would thus not be communicatively functional, although the underlying strategy of including salient properties is a manifestation of cooperative behaviour.

The present studies were conducted to explore patterns in colour overspecification that are strictly non-functional, but due to the more general, cooperative strategy to include salient attributes. We do not test the effect of colour salience directly, but we investigate how colour overspecification of objects in various categories is affected by factors which, presumably, contribute to colour salience of those objects. Study 1 is a production experiment conducted to assess whether the tendency to produce colour overspecification is affected by the degree to which colour is *important* (and hence probably salient) for the referent. We compare references to different types of referents: clothes (high colour importance), dinner ware (medium colour importance), and office supplies (low colour importance). A pretest was conducted to establish subjective ratings of colour importance. We hypothesise that the likelihood of colour overspecification increases with colour importance of the referent.

In Study 2, we investigate colour overspecification in reference to a special category of referents: geometrical figures. It is fairly common to investigate referential behaviour experimentally by making participants refer to geometrical figures (Mangold and Pobel, 1988; Arts et al., 2011b). Geometrical figures are easy to manipulate, but they are abstractions rather than real objects. As they have no other attributes than shape and colour to attract the attention, their colour might be more salient than the colour of real life objects whose colour is equally important. We hypothesise that this *paucity* of attributes that may attract the attention is a second factor in colour salience and hence in the production of colour overspecification. In Study 2, we investigate colour overspecification in reference to figures, comparing this category with a category of objects whose colour is equally important.

2 Pretest

In order to be able to select the items for Studies 1 and 2, we conducted a pretest to assess to what extent speakers of Dutch judged colour to be important for objects in various categories. To this end, we presented participants with pictures of objects and asked them to judge, on a 7-point scale, how important they felt colour was for the object in question. This procedure enabled us to select four objects in four categories: one with high, one with medium, and two with low colour importance.

2.1 Method

2.1.1 Participants

We tested 21 native speakers of Dutch (18 females, 3 males, mean age 22:1 years, range 18-27) at Radboud University, the Netherlands. All were volunteers. They received a small fee for their participation.

2.1.2 Materials

We used 60 black-and-white photographs of objects as stimulus materials, divided into ten categories (objects to draw, write, or paint with, clothes, vehicles, toys, dinner ware, furniture, kitchen utensils, office supplies, cleaning utensils, and geometrical figures) of six objects each. All real life objects were familiar items which are commonly available in a variety of colours and which are easily recognised and named. Additionally, three filler items were included, which did not belong to any of the ten categories.

Each photograph represented one object against a white background. The selection criteria were that the object should be easy to recognise and that the photograph should be as simple as possible. The original photographs were freely available on the internet. Some were manipulated in Photoshop. Only photos of painted objects were selected, in order to avoid an association with the typical colour of certain materials (such as unpainted wood, which is typically brown). This experiment and all the following experiments were programmed with Presentation software.

2.1.3 Design

All participants judged the colour importance of each of the 60 items. The order of the items was pseudorandomised, with the restriction that items were always followed by at least two items from a different category. Each participant saw the items in a different order.

2.1.4 Procedure

Participants were tested one at a time in a quiet booth. In each trial, participants saw a picture of an object and a 7-point scale below it on a computer screen. Participants were instructed to indicate, by clicking on a point on the scale, how important they felt colour was for the object, where 1 represented ‘not at all important’, and 7 ‘highly important’. They were encouraged to follow their intuitions and react quickly. There was no timeout for responding. It took participants about five minutes to complete the task.

2.2 Results and selection procedure

We excluded one of the ten categories[1] from further consideration. For the remaining nine categories, the median judgements of colour importance of the items are represented in Figure 1. We selected those items which we expected to be easy to recognise and name for speakers of Dutch, and that were not visually or conceptually similar to another item in the same category (such as a circle and an ellipse). We selected categories with four items that were as homogeneous as possible in their median judgement.

For Study 1, we selected a High Importance category (*Mdn* = 6), a Medium Importance category (*Mdn* = 4), and a Low Importance category (*Mdn* = 2). We selected clothes as High Importance category (trousers, coat, dress, all *Mdn* = 6, and hat, *Mdn* = 5), dinner ware as Medium Importance category (plate, mug, bowl, all *Mdn* = 4, and teapot, *Mdn* = 3), and office supplies as Low Importance category (stapler, pencil sharpener, scissors, all *Mdn* = 2, and ring binder[2], *Mdn* = 3). For Study 2, we selected four geometrical figures (circle, square, triangle, diamond, all *Mdn* = 2).

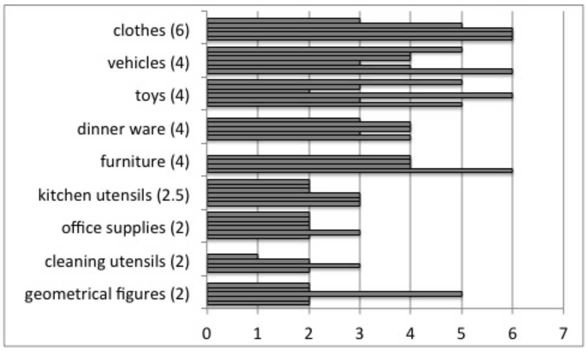


Figure 1: Median judgements of colour importance for the items in each category. The integer in brackets behind the category labels represents the median for that category.

3 Study 1: Colour importance

In Study 1, we tested the hypothesis that there is a positive correlation between judgements of colour importance and the amount of colour overspecification, by conducting a production experiment in which participants referred to objects of the three categories of real life objects selected in the pretest.

3.1 Method

3.1.1 Participants

We tested 38 participants similar to those in the pretest (33 females, 5 males, mean age 22:10 years, range 18-29). None of the participants in Study 1 had taken part in the pretest. All of them reported not to be colourblind.

3.1.2 Materials

Twelve critical pictures represented the objects selected in the pretest. They were found on the internet and then manipulated in Photoshop to create four colour variants of each picture: bright red, green, yellow, and blue.[3] This procedure thus yielded 48 different pictures altogether. We

---

[1] The category of objects to draw, write, or paint with was excluded because expressions such as ‘the green pen’ are ambiguous between a pen filled with green ink and a pen painted green.

[2] Although there was a fourth object with a median of 2, namely the paperclip, we selected the ring binder instead because it was impossible to sufficiently increase the coloured area of the paperclip picture (see section 3.1.2 below).

[3] The pictures in the experiment and the pretest were as similar as possible. We did not use the pictures from the pretest because most of them were not suitable for making good colour variants.

constructed the pictures so that the size of the coloured area was approximately similar across categories[4] (mean number of coloured pixels per picture: 28505 for clothes, 29821 for dinner ware, and 29703 for office supplies).

Filler pictures were taken from the Tarrlab Stimulus Repository[5]. There were three types of filler pictures: sixteen common objects such as bikes and envelopes (Rossion and Pourtois, 2004), sixteen Greebles (Gauthier and Tarr, 1997), and sixteen human faces. Greebles are artificially constructed objects which are complex and highly similar to each other, and therefore difficult to describe uniquely. Paying attention to colour was prevented by changing salient colours into desaturated, inconspicuous ones (common objects) or into tones of grey (Greebles), and by selecting pictures of dark-haired Caucasian people only (human faces).

3.1.3 Design

Participants were randomly assigned to one of three conditions: High Importance, Medium Importance, and Low Importance. Colour importance was manipulated between participants: each participant saw objects from only one of the three categories. Each of the four objects in a category acted as target four times (in four different colours), so that each participant performed sixteen critical trials. They also performed sixteen trials of each of the three types of fillers, yielding a total of 64 trials. The order of the trials was pseudorandomised, with the restriction that each trial was always followed by at least two trials in which the target was of a different type of object. For example, when the target was a dress, the target in the two subsequent trials was never a dress. This was done to prevent participants from producing an adjective for the sake of contrast between the referent and the previous referent. Each participant received the trials in their own unique order.

Target pictures were presented in an array with other objects of the same category. The number of items in an array varied among two, three, four, and six. The objects in the context were never of the same type as the target object. Including colour therefore always resulted in colour overspecification, except for the rare cases where participants did not use a basic-level term (e.g., ‘the yellow object’ instead of ‘the yellow stapler’), which were not included in the analysis. Colours were pseudorandomly distributed over the objects in the array, with the restriction that monochrome displays did not occur. The target could be, but was not necessarily unique in its colour.

Fillers were added to prevent participants from sticking to one syntactic and semantic structure throughout the whole experiment, and from finding out about the aim of the experiment. There were three types of filler trials. Fillers of type A were displays with four pictures of common objects. They were included to elicit referring expressions in which no modifier, such as an adjective or a prepositional phrase, was added to the head noun. Modification was not expected because basic-level terms were always sufficient and none of the pictures had any striking features. Fillers of type B were displays with four pictures of Greebles. They were included to make participants aware that simply naming objects was not always sufficient. Fillers of type C were displays with two human faces, which were either of the same gender or of different genders. They were included to elicit variation in the presence of modifiers within a category: modification was necessary when the two people were of the same gender, but unnecessary when they differed in gender.

3.1.4 Procedure

Participants had to instruct an imaginary addressee to click on one of the pictures displayed in each trial, by finishing the Dutch equivalent of the sentence ‘Click on . . . ’. A cross preceding the presentation of the array indicated the position of the target on the screen. Participants were instructed to avoid referring to the object’s location on the screen. It took them about fifteen minutes to complete the task. Otherwise the procedure was similar to that of the pretest.

3.2 Results

Each of the 38 participants performed sixteen critical trials, yielding 608 responses. Twenty responses (3.3%) were removed, because the referent was not the target item, because the speaker corrected herself during the articulation of the utterance, or because colour was included without

---

[4] The results of a pilot study made us suspect a positive correlation between the size of the coloured area of a picture and the probability of colour overspecification.

[5] Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/. For some of the pictures we adjusted the colours or we flipped them into a mirror image.

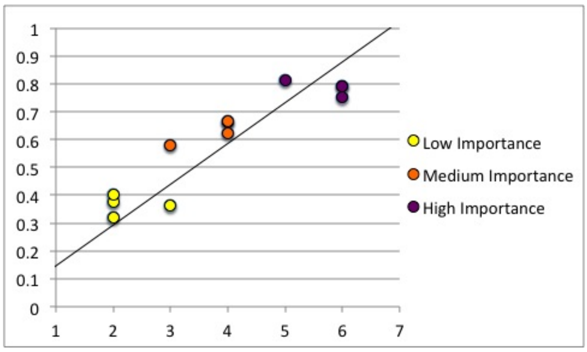


Figure 2: The relation between colour importance and colour overspecification. The median colour importance ratings are plotted on the x-axis, and the mean proportions of colour overspecification are plotted on the y-axis.

this resulting in overspecification. The remaining 588 expressions were annotated as colour overspecified when a colour adjective was included.

We expected that the proportion of colour overspecification would increase with the degree to which colour is considered important for the object. That is, we expected a positive correlation between the colour importance judgements collected in the pretest, and the mean proportions of colour overspecification produced in reference to those items in the present experiment. Indeed, as Figure 2 shows, the proportion of colour overspecification increased with colour importance. The mean proportion of colour overspecification was highest in the High Importance condition ($M = .79, SD = .41$), intermediate in the Medium Importance condition (M = .63, SD = .48), and lowest in the Low Importance condition (M = .37, SD = .48). The correlation between the median judgements of colour importance and the proportions of colour overspecification of the items was significant, τ = .762, 95% CI [.335, .929], p = .001.

3.3 Discussion

We predicted that the salience of an object's colour would increase with the degree to which colour is considered important for that object, resulting in a higher proportion of colour overspecification in reference to the object. Our prediction was borne out by the results: there was a significant positive correlation between colour importance judgements and the mean proportion of colour overspecification in reference to the same items.

Since the pretest indicates that colour importance is considered to be equally low for geometrical figures as for office supplies, speakers are not expected to often produce colour overspecification when referring to figures. However, as pointed out in the Introduction, colour salience is probably not only determined by colour importance, but also by the number of other attributes that matter: if only a low number of attributes may attract the attention, those attributes will increase in salience. The colour of simple geometrical figures might be highly salient because the only attributes of geometrical figures that matter are colour and shape. This possibility was investigated in Study 2.

4 Study 2: Geometrical figures

Study 2 was conducted to test the hypothesis that speakers produce more colour overspecification when referring to geometrical figures than to objects of equal colour importance. We elicited references to figures and compared the amount of colour overspecification to the amount produced in Study 1 in reference to office supplies, as the Pretest had indicated that colour is considered to be equally important for the two categories.

4.1 Method

4.1.1 Participants

We tested 13 participants similar to the ones in Study 1 (all females, mean age 21:3, range 19-26).[6] None of the participants in Study 2 had participated in either of the previous studies.

4.1.2 Materials, design, and procedure

Critical pictures represented the geometrical figures selected in the pretest. They were created in LaTeX, using the Tikz package, sometimes in combination with Photoshop. Otherwise, materials, design, and procedure were as in Study 1.

4.2 Results

Each of the 13 participants performed 16 critical trials, yielding 208 responses, 23 (11%) of which were removed as in Study 1. The remaining 185 expressions were annotated as in Study 1.

The experiment was conducted to test the hypothesis that speakers produce more colour overspecification when referring to geometrical figures than to office supplies. To this end, we compared the proportion of colour overspecification produced in Study 2 to that produced in the Low Importance condition (office supplies) in Study 1.

---

[6] Two additional participants participated in the experiment but their data were not analysed, because colour was included without resulting in overspecification in more than half of the trials (n = 1) or because they did not understand the task (n = 1).

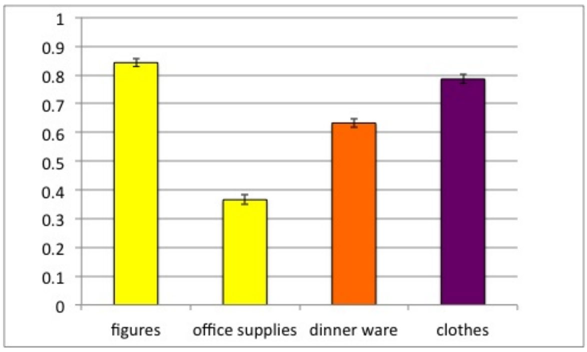


Figure 3: Mean proportions of colour overspecification for geometrical figures from Study 2, and office supplies (Low), dinner ware (Medium), and clothes (High), from Study 1. The error bars represent standard errors.

Figure 3 represents the mean proportions of colour overspecification in reference to geometrical figures and office supplies. For reasons of comparison, the mean proportions for dinner ware (Medium Importance) and clothes (High Importance) from Study 1 are also represented. As hypothesised, the proportion of colour overspecification was higher in reference to geometrical figures (*M* = .84, *SD* = .37) than in reference to office supplies (*M* = .37, *SD* = .48). The individual participants' proportions of colour overspecification varied a lot within conditions, as the high standard deviations suggest. A Shapiro-Wilk test indicated that the data were not normally distributed (*p* was below .05 in both conditions). We therefore ranked the data (we report mean ranks, denoted by *MR*) and used non-parametric statistics. A Mann-Whitney test indicated that the difference between geometrical figures (*MR* = 16.58) and office supplies (*MR* = 9.12) was significant and that the effect size was large, *U* = 31.50, *z* = -2.59, *p* = .01, *r* = -*.52*.

As can be seen in Figure 3, the proportions of colour overspecification produced in reference to geometrical figures and clothes (the High Importance condition in Study 1) were very close. A Mann-Whitney test indicated that the difference between figures (*MR* = 12.27) and clothes (*MR* = 13.79) was not significant, *U* = 87.50, *z* = .60, *p* = .61, *r* = .12.

4.3 Discussion

Study 2 was conducted to test the hypothesis that speakers are more likely to produce colour overspecification in reference to geometrical figures than to office supplies, even though colour is of equally low importance for the two categories. This prediction was borne out by the data. In fact, the proportion of colour overspecification produced in Study 2 was so high, that is was statistically indistinguishable from the proportion produced in reference to clothes, the High Importance condition in Study 1. The results suggest that the colour of geometrical figures is substantially more salient than the colour of office supplies, which we have argued to be due to the fact that geometrical figures are very simple objects whose only attributes which may attract the attention are colour and shape.

5 General discussion and conclusions

We presented a series of experimental studies that investigate the production of colour overspecification in reference to objects in different object categories. In Study 1, we tested the hypothesis that salience of the colour of objects, and hence the probability that speakers produce colour overspecification when referring to those objects, increases with the degree to which colour is considered important for objects. In this experiment, participants referred to objects that we know from a pretest to vary in colour importance: clothes (High Importance), dinner ware (Medium Importance), and office supplies (Low Importance). We found a significant positive correlation between the median ratings of colour importance and the mean proportions of colour overspecification, which is evidence for our hypothesis.

The pretest indicated that colour is considered about equally important for geometrical figures as for office supplies. In Study 2, we investigated whether objects in the two categories nevertheless diverge in how likely speakers are to produce colour overspecification when referring to them. We predicted that the colour of simple geometrical figures is more salient than the colour of office supplies because figures have a low number of attributes that may attract the attention, and that speakers are hence more likely to produce colour overspecification when referring to figures than to office supplies. This prediction was corroborated by the data, which is in line with previous studies in which high rates of colour overspecification were found in reference to geometrical figures (Arts et al., 2011b). Besides, speakers referring to figures produced a very similar amount of colour overspecification to speakers who referred to clothes, to which colour is highly important.

We conclude from Studies 1 and 2 that the likelihood of colour overspecification increases when colour is relevant to the referent, and when the referent has a low number of attributes that may attract the attention. We have argued that colour relevance and paucity of attributes both increase colour salience, which triggers selection of colour, even if the resulting colour adjective is redundant.

It might be questioned whether colour importance really increases the salience of an object's colour, as this hypothesis was tested only indirectly. An alternative explanation is that the colour of office supplies is equally salient to the colour of clothes, but that some speakers do not select colour when they are referring to office supplies because the lack of colour importance makes them realise that colour is redundant. We think this unlikely, because out of the seven participants in the Low Important condition in Study 1 who produced colour overspecification at least once, six had not produced it in the first trial, and four kept producing it consistently after the first time they did include colour. That is, if they had realised that colour was redundant in their first trial, why then would they start to include it later in the experiment? We therefore maintain that it is salience of an object's colour that largely determines whether colour will be included in a referring expression. This is not to say that a high degree of salience of an attribute automatically leads to including it. It is perfectly possible, and indeed likely, that speakers evaluate to some degree whether a selected attribute is sufficiently important. However, the fact that colour overspecification is sometimes produced in monochrome contexts suggests that such an evaluation mechanism is not infallible.

The question remains, however, why colour importance would increase colour salience. A possible answer to this question is that when colour is important to an object, speakers will often include the colour of such an object when talking about it even in situations where the intention is not to enable the addressee to identify a referent, but rather to feed his imagination such that he can shape an accurate image of the object in his mind. For example, Bill may tell Ann-Marie about his beautiful new pink shirt, without intending to enable her to pick out the right object as a referent, but just to give her an idea of what his precious purchase looks like. If colour is important to an object, people may therefore be inclined to pay attention to it. Moreover, as the label of an object is often accompanied by a colour term, an association may emerge between the colour term and this label.

As was argued in the Introduction, we claim that the effect of object categories on how likely speakers are to produce colour overspecification is due to a general cooperative strategy: selecting salient attributes generally leads to efficient identification of the referent. We think it unlikely that speakers tend to produce colour overspecification in reference to clothes but not to office supplies because they reckon their addressee will benefit from colour in identifying clothes but not in searching for office supplies. Only empirical evidence can tell us whether colour is more beneficial in identifying clothes than office supplies. As was pointed out in the Introduction, overspecification has been found to be beneficial in some studies but cumbersome in others, and why experimental results diverge at this point is as yet unclear. Addressees may be more likely to notice that colour is redundant when the referent is a stapler than when it is a dress, and hence are hindered by colour overspecification in the former case but not in the latter. Our point is that whether or not this is the case, it is not the reason why speakers select colour more often when referring to dresses than to staplers.

6 Conclusions

A series of production experiments showed that speakers are more likely to produce colour overspecification when referring to some objects than to others, apparently regardless of how helpful colour is for identifying the objects. Colour overspecification increased with colour importance in reference to real life objects. It was also high in reference to geometrical figures, even though colour importance is low for this category. We argue that colour overspecification increases with colour salience, and that colour importance of real life objects and a paucity of attributes that may attract attention both contribute to colour salience. We claim that this is due to a general cooperative strategy, because in general, salient attributes are likely to be helpful in the identification process.

Acknowledgements

We thank Ronald Fischer and Frauke Hellwig for their help in programming the experiments, Bob van Tiel for testing participants, and three anonymous reviewers for their comments and questions. This research was supported by a grant from the Netherlands Organisation for Scientific Research (NWO), which is gratefully acknowledged.

References

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Anja Arts, Alfons Maes, Leo Noordman, and Carel Jansen. 2011a. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

Anja Arts, Alfons Maes, Leo Noordman, and Carel Jansen. 2011b. Overspecification in written instruction. *Linguistics*, 49(3):555–574.

Eva Belke and Antje S. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, 14(2):237–266.

Eva Belke. 2006. Visual determinants of preferred adjective order. *Visual Cognition*, 14(3):261–294.

Brent Berlin and Paul Kay. 1969. *Basic color terms: Their universality and evolution*. University of California Press, Berkeley.

Sarah Brown-Schmidt and Agnieszka E. Konopka. 2011. Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversations. *Information*, 2(2):302–326.

Catherine Davies and Napoleon Katsos. 2013. Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics*, 49(1):78–106.

Paul E. Engelhardt, Karl G.D. Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54(4):554–573.

Paul E. Engelhardt, Ş. Barış Demiral, and Fernanda Ferreira. 2011. Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, 77(2):304–314.

William Ford and David Olson. 1975. The elaboration of the noun phrase in children's descriptions of objects. *Journal of Experimental Child Psychology*, 19(3):371–382.

Albert Gatt, Emiel Krahmer, Roger P.G. van Gompel, and Kees van Deemter. 2013. Production of referring expressions: Preference trumps discrimination. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 483–488.

Isabel Gauthier and Michael J. Tarr. 1997. Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision research*, 37(12):1673–1682.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics, Vol. 3: Speech acts*, pages 41–58. Academic Press, New York.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Ruud Koolen, Martijn Goudbeek, and Emiel Krahmer. 2013. The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37(2):395–411.

Roland Mangold and Rupert Pobel. 1988. Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology*, 7(3–4):181–191.

Thomas Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27(1):89–110.

Bruno Rossion and Gilles Pourtois. 2004. Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2):217–236.

Paula Rubio-Fernández. 2011. Colours and Colores. Plenary talk at the 4th Biennial Conference on Experimental Pragmatics, Universitat Pompeu Fabra, Barcelona.

Julie C. Sedivy. 2005. Evaluating explanations for referential context effects: Evidence for Gricean mechanisms in online language interpretation. In John C. Trueswell and Michael K. Tanenhaus, editors, *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pages 345–364. MIT Press, Cambridge, Massachusetts.

Anne M. Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Kees van Deemter, Albert Gatt, Roger P.G. van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in cognitive science*, 4(2):166–183.

Jette Viethen, Robert Dale, and Markus Guhe. 2014. Referring in dialogue: alignment or construction? *Language, Cognition and Neuroscience*, 29(8):950–974.

Hans Westerbeek, Ruud Koolen, and Alfons Maes. 2014. On the role of object knowledge in reference production: Effects of color typicality on content determination. In *CogSci 2014: Cognitive science meets artificial intelligence: Human and artificial agents in interactive contexts*, pages 1772–1777.

Editing Phrases

Ye Tian[1], **Claire Beyssade**[2], **Yannick Mathieu**[1], **Jonathan Ginzburg**[3,4]
[1]Laboratoire Linguistique Formelle (UMR 7110) & [2]Institut Jean Nicod (UMR 8129)
& [3]CLILLAC-ARP (EA 3967) & [4]Laboratoire d'Excellence (LabEx)—EFL
Université Paris-Diderot, Paris, France
`tiany.03@googlemail.com`

Abstract

Disfluencies are viewed as a performance phenomenon in most formal grammatical treatments. In this paper we provide evidence for the need to integrate disfluencies in the competence grammar. We do this by considering the properties of editing phrases (EPs). We study their distribution in the American English corpus Switchboard and the French corpus Rhapsodie. We show that English and French exhibit various distributional differences, as expected from a grammatical phenomenon. We sketch a treatment for distinct classes of editing phrases.

1 Introduction

Disfluencies are viewed as a performance phenomenon in most formal grammatical treatments, though this view is explicitly rejected by psycholinguists e.g., (Levelt, 1983; Clark and FoxTree, 2002) and some theoretical linguists (Blanche-Benveniste, 1984; De Fornel and Marandin, 1996; Ginzburg et al., 2014; Husband, 2015). In this paper we provide evidence for the need to integrate disfluencies in the competence grammar. We do this by considering the properties of editing phrases (EPs).

As a terminological preliminary, we adopt Jens Allwood's term 'own communication management' (OCM) instead of 'disfluency' (Allwood et al., 2005). When a speaker interrupts her utterance with an OCM element, she often uses an editing phrase (EP) to signal a correction or reformulation. A typical structure of self-repair can be illustrated by figure 1, annotated with the labels introduced by (Shriberg, 1994), who was building on (Levelt, 1983).

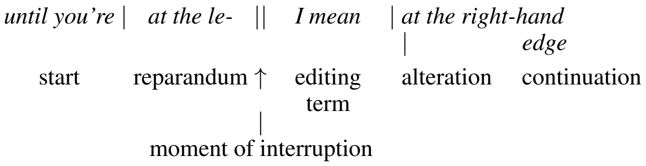


Figure 1: General pattern of self-repair

To determine whether a word or phrase is an editing phrase, one can resort to its semantic meaning, the structural context or both. The annotation guideline for Switchboard defined editing terms as "having some semantic content, e.g. I mean, sorry, excuse me) and usually occur between the restart and the repair". This definition primarily uses the semantic meaning to determine a term's potential of being an EP. If a pause filler (e.g. 'uh') is used between a reparandum and a repair, it will not be categorized as an EP. On the other hand, one could define an EP using just its structural context (e.g., (Levelt, 1983)), which is the approach we adopt here.

Although some differences in use among EPs have been noted in earlier work (see e.g., (Tree and Schrock, 2002)), the substantial syntactic and semantic differences among them have not been detailed. Thus, some EPs can participate in backwards looking (BL) (corrective) OCMs, but not in forwards looking (FL) (monotonically continuative) OCMs:

(1) a. Bo is forty excuse me / or / no fifty.

b. Bo is um # excuse me / or / no fifty.

c. Bo is you know / like fifty.

Similarly, some EPs can occur turn initially, but many cannot:

(2) a. A: Where did you leave the book?

B: # No / Or . . . in the bathroom.

b. uh / you know / I don't know . . . in the bathroom.

The main aims of the paper are these. We attempt to demonstrate that editing phrases

- exhibit properties that require stating in an interaction oriented grammar
- exhibit cross-linguistic variation and systematic behaviour

We start in section 2 by describing the distribution of EPs in the American English corpus Switchboard. Section 3 describes the distribution of EPs in the French corpus Rhapsodie. In section 4 we offer some comparative discussion. In section 5 we sketch a formal account of the difference between backwards and forwards looking EPs. Section 6 contains some brief conclusions and future work.

2 The Distribution of Editing phrases in Switchboard

We searched in the Switchboard Dialogue Act Corpus (SWB) (Stolcke et al., 2000)) for OCMs with the structure annotation of [reparandum, +{editing phrase} repair]. This returns OCMs that are repeats (see example 3b) or revisions (example 3a):

(3) a. ('you know' BL) we don't, you know, I don't ask for more.(sw_0049_4353)

b. ('you know' FL) Because I've caught up to about an eight pound carp on a little, you know, a little pole with twenty pound test line. (sw_0563_3458)

The former is an example of backward-looking OCM and the latter forwards-looking.

In Table 1, we list the editing phrases (EP), their number of occurrences as EP (totaling 1942 cases) and the ratios of repeat to revision.

expr.	EP freq.	B/F	total freq.	% post reperandum	ratio repeat: revision
you know	1216	BF	13940	8.72%	1.78:1
well	160	BF	9453	1.68%	1.14:1
I mean	183	BF	3470	5.27%	0.8:1
or	101	B	172	58.72%	0:1
like	95	BF	1970	4.82%	1.07:1
yeah	62	BF	231	26.84%	11:1
oh	49	BF	7348	0.67%	1.18:1
actually	24	BF	314	7.64%	0.71:1
no	5	B	12	33.33%	0:1
excuse me	4	B	45	8.89%	0:1

Table 1: Editing phrases in Switchboard

In English, EPs seem for the most part to perform both BL and FL functions:

(4) a. ('well' BL) but I do live in the better, well, in the best part of the city. (sw_0064_4346)

b. ('well' FL) You, + you'd, well, you'd think there would be. (sw_0392_2405)

c. ('I mean' BL) Well that would be sort of interesting because then you get people from other countries, I mean other parts of the state you know.(sw_0737_2110)

d. ('I mean' FL) I wonder, I mean, I wonder what what really is the answer.(sw_0046_4316)

e. ('like' BL) We're still, like, I'm still covered under my mom and dad's life insurance because I'm still in school. (sw_0998_2175)

f. ('like' FL) Now, do you usually, like, do you usually go and there's lots of other people around (sw_1003_2524)

g. ('yeah' BL) Whatever's left over is disposable in-, disposable, dis-, yeah, discretionary income.(sw_0631_4149)

h. ('yeah' FL) I bet that was a good day, at the, yeah, conference then. (sw_0027_4096)

i. ('oh' BL) She says that when her husband died that he said, oh, that my uncle had said that he would never ha- put her in a rest home. (sw_0351_3207)

j. (‘oh’ FL) we take, oh, one big vacation a year and then maybe, you know, three small vacations.(sw_0036_4379)

However, there are some EPs that can only functions as BL:

(5) a. (‘or’ BL) Myself, uh, uh, I’m just recently, or about to get a divorce. (sw_0097_3798)

b. (‘actually’ BL) I have a foreign, actually I have more than one foreign automobile. (sw_0932_2610)

c. (‘excuse me’ BL) a table saw does take a lot of time, excuse me, a lot of space and is a pretty big investment. (sw_0627_3651)

d. (‘no’ BL) they have one of the clerks up there, no, the bag boys out there, um, that will take the papers, newspaper out of your car.(sw_0798_3736)

Note that in SWB there does not seem to be an editing phrase that is solely FL. Is this a deep fact about the grammar or an accidental feature of this corpus study? ‘I don’t know’ is a candidate to be exculsively FL, but it isn’t a pure one. On the one hand, it resists parallelism repairs, as in (6c), but allows fresh starts, as in (6d); it also can occur on the right periphery of an utterance, as in (6e):

(6) a. So if somebody, in I dont know, Penge (South London) said I could deliver you 50 votes you would laugh. (Ben Judah, Politico, May 6, 2015)

b. I’ve only got eight more things to get her, I’ve already spent about, I don’t know, sixty quid on her. (BNC, KDA)

c. # So if Bill, I dont know, Mary said …

d. Unknown: It’s er, I don’t know, they’re having a ⟨unclear⟩at us ⟨unclear⟩(BNC, KPL)

e. And erm so, of course, the land army came in then and erm 1939, September, I dont know, there were 900 volunteers already. (BNC, KRX).

expr.	EP freq.	B/F	total freq.	% post reperandum	ratio repeat: revision
uh	2864	BF	25391	11.29%	1.92:1
um	376	BF	6482	5.77%	2.36:1

Table 2: FIllers in Switchboard

We found 3289 cases where pause fillers ‘uh’ and ‘um’ were used as an editing phrase (in the post-reparandum position in a repair).

We also found around 23000 instances of OCM of the structure [reparandum, repair] (where no editing terms were inserted), and the ratio of repeats to revision is 3.6:1.

The data we have seen in this section suggests that:

1. In English, the majority of the time, speakers do not use an editing phrase in self-repair. (about 2000 revision OCMs with an editing phrase, and about 5000 without). This pattern is not a linguistic universal. Levelt (1983) shows that in Dutch, self-repairs with EPs make up over half of all self-repairs.

2. A relatively wide range of words can be used as EPs. The semantic meaning of an EP does not always suggest the correction of the reparandum (e.g. ‘you know’ and ‘like’).

3. However, some EPs are more corrective than others. ‘Or’, ‘no’ and ‘excuse me’ can only be used to revise. ‘I mean’ and ‘actually’ are used more often in revisions than in repeats. All of these terms have in their semantic content the element of correction.

4. When an EP is inserted, the OCM is more likely to be revision than repeat (with the exception of ‘yeah’). The highest revision to repeat ratio for EPs is ‘um’ at 2.36:1, which is still considerably lower than the ratio of EP-less repairs (3.6:1).

(7) a. (‘uh’ BL) I think uh I wonder if that worked. (sw_0001_4325)

b. (‘uh’ FL) Well, we’ve always, uh, we’ve always had Oldsmobiles, and, uh, been very, uh, happy with Oldsmobiles. (sw_0191_3427)

c. ('um' BL) Sometimes, um, usually the reason I will turn it on is to hear the news. (sw_0249_3728)

d. ('um' FL) And it's been, um, and it's been pretty rainy (sw_1044_2457)

3 The Distribution of Editing phrases in Rhapsodie

(Pallaud et al., 2013) (note 7 page 23) propose a list of the editing phrases in French, based on their research in the CID corpus (Bertrand et al., 2008), though offer no data concerning distribution:

(8) ah, ben, bof, bon, bref, daccord, eh, enfin, euh, hein, jen sais rien, je sais pas, là, oh, non, ouais, oui, putain, quoi, si tu veux, tu vois, tu sais, voilà.

We used this as a basis for searching the Rhapsodie corpus (Lacheret et al., 2014), which is annotated for OCMs. As (Gerdes et al., 2012) explain: 'The corpus is made up of 57 samples of spoken French (5 minutes on average) mainly drawn from existing corpora of spoken French for a total of 3 hours and 33 000 words and distributed under a Creative Commons licence at `http://www.projet-rhapsodie.fr`.'[1]

The results are provided in table 3:

expression	post repr	Back/For	total freq	% post repr
Euh	933	B/F	1008	92%
Hein	73	B/F	87	84%
Enfin	60	B/F	81	74%
Oui	42	F	244	17%
Non	20	B	155	13%
Eh	23	?	33	70%,
Ouais	16	F	88	18%
Quoi	13	F	48	27%
Voilà'	13	F	72	18%
Disons	12	B/F	17	70%
Je sais pas	2	F	13	12%

Table 3: Editing phrases and fillers in Rhapsodie

Apart from the filler 'euh' and 'hein', the commonest EP is 'enfin' (lit. 'finally'), which is both BL and FL:

(9) a. Je connaissais très bien Marc Allégret depuis très longtemps. **Enfin** ma famille le connaissait . (*I knew very well Marc Allégret since a long time. Enfin my family him knew.*)

b. Euh il faut également proposer des méthodes pédagogiques qui — qui visent à intéresser euh donc son auditoire — donc sa classe euh pour parvenir à de très bons résultats // et euh également c'est le fait euh **enfin** de corriger mais également de réaliser des devoirs (*um it is necessary equally to propose methods pedagogical that that target to interest um so his listener so his class um to manage very good results and um equally the fact um enfin to correct but equally to realize the homework*).

Another EP which is both BL and FL is *disons* (lit. say-1pl). *Disons* is not exactly corrective, but it is used to reformulate with more appropriate words. It can precede (cf (10a)) or follow the reformulation (cf (10b)):

(10) a. Nous étions tous les deux d'origine bourgeoise, élevés un peu de la même manière, euh c'est-à-dire **disons** d'une façon un peu britannique dans le comportement, n'est-ce pas (*We were all three of origin bourgeois, students a little in the same manner um that is to say disons in a manner a bit British in the behaviour*)

b. Mais j'ai tendance à — à penser par phrases **disons** — et pas à penser par pensées. (*But I had the tendency to to think by phrases disons by thoughts.*)

There are some variants of *disons*, like *si je puis dire* (lit. *if I could say*) or *on peut dire, on va dire* (lit. *one could say, one will say*). They seem to be only forward looking, never corrective:

(11) a. Est-ce que vous vous êtes fixé un cadre, **si je puis dire**, dans la durée? (*Have you you fixed a framework si je puis dire in the duration*)

---

[1] An anonymous reviewer for SemDial 2015 wonders whether differences between the nature of SWB—two person phone conversations between strangers—and Rhapsodie might lead to distributional differences. This is an interesting question which we hope to be able to address in future work using, on the one hand, a larger French corpus we are currently compiling. On the other hand, also using the British National Corpus, which is more balanced than SWB, though involves British English, which involves distinct distributions of OCMs than SWB.

b. Est-ce que vous vous êtes fixé un cadre, **on va dire**, dans la durée? (*Have you you fixed a framework on va dire in the duration*)

Similarly for 'voilà' (lit. 'there') in Rhapsodie. There are no examples of 'voilà' with a corrective value.

(12) a. Donc on a beaucoup de mal à maintenir **voilà** une clientèle de quartier (*So one had a lot of trouble to maintain voilà a clientèle in the neighbourhood'*)

b. J'ai pas été acceptée parce qu'il y avait un entretien oral et je le savais pas donc en fait euh **voilà** c'est, c'est trop enfin stupide. (*I was not accepted because there was a conversation oral and I didn't know so in fact um voilà that's that's too enfin stupid.*)

As with its English counterpart, *Je sais pas* seems inappropriate to introduce a correction:

(13) a. je suis pas du tout une acharnée de — de l'actualité littéraire // et je suis euh **je sais pas** quoi épouvantablement éclectique quoi (*I am not at all a devotee of the goings on literary and I am um je sais pas somewhat dreadfully eclectic like.*)

b. # je connaissais très bien Marc Allégret depuis très longtemps. Je sais pas ma famille le connaissait .

Whereas *non* is also like its English counterpart in being solely BL:

(14) a. ah moi je suis une fille extrêmement pudique dans le fond, **non** mais même pas dans le fond, je suis très, très pudique. (*Ah, me I am a girl extremely modest at the bottom, No but even not at the bottom I am very very modest.*)

b. et le ballon est sorti pour euh l'équipe de France là, **non** pour les Argentins (*And the ball is out for um team France there non for the Argentinians.*)

4 Crosslinguistic differences/commonalities between Editing phrases and their implications

The commonest non–filler EPs in Switchboard are, by some distance, 'you know', 'well', and 'I mean'; the commonest French ones in Rhapsodie are 'hein', 'enfin', and 'oui'. Though 'you know and ''hein' correspond roughly—they both have uses to make *check* moves—this is a strong illustration that the distribution of EPs is highly language–specific.

At the same time, there is evidence that certain semantic properties of EPs are preserved under translation:

- (Ginzburg et al., 2014) propose a universal concerning negative EPs—*if NEG is a language's word that can be used as a negation and in cross-turn correction, then NEG can be used as an editing phrase in backward-looking OCMs.* ((Ginzburg et al., 2014), p.10). English 'no' and French 'Non' can indeed both serve as EPs and both are only BL EPs
- Conversely English 'I don't know' and French 'Je sais pas' are both FL EPs

5 Editing phrases: formal analysis

5.1 Background

We rely on the approach to OCMs developed by (Ginzburg et al., 2014) using the dialogue framework KoS (see e.g., (Ginzburg, 2012) for details). The dialogue gameboard represents the public part of a participant's information state. Its structure is given in (15)—the *spkr,addr* fields allow one to track turn ownership, *Facts* represents conversationally shared assumptions, *Pending* and *Moves* represent respectively moves that are in the process of/have been grounded, *QUD* tracks the questions currently under discussion.

(15)
$$\text{DGBType} =_{def} \begin{bmatrix} \text{spkr: Ind} \\ \text{addr: Ind} \\ \text{utt-time : Time} \\ \text{c-utt : addressing(spkr,addr,utt-time)} \\ \text{Facts : Set(Proposition)} \\ \text{Pending : list(locutionary Proposition)} \\ \text{Moves : list(locutionary Proposition)} \\ \text{QUD : poset(Infostruc)} \end{bmatrix}$$

Metacommunicative interaction is handled in KoS by assuming that in the aftermath of an utterance u it is initially represented in the DGB by means of a *locutionary proposition* individuated by u and a grammatical type T_u associated with u. If T_u fully classifies u, u gets grounded, otherwise clarification interaction ensues regulated by a question inferable from u and T_u. If this interaction is successful, this leads to a new, more detailed (or corrected) representation of either u or T_u.

(Ginzburg et al., 2014) develop their account in KoS of OCMs by extending the account just mentioned of the coherence and realization of clarification requests: as the utterance unfolds incrementally there potentially arise questions about what has happened so far (e.g. *what did the speaker mean with sub-utterance u1*?) or what is still to come (e.g. *what word does the speaker mean to utter after sub-utterance u2*?). These can be accommodated into the context if either uncertainty about the correctness of a sub-utterance arises or the speaker has planning or realizational problems. Thus, the monitoring and update/clarification cycle is modified to happen at the end of each word utterance event, and in case of the need for repair, a repair question gets accommodated into QUD.

5.2 Distinguishing distinct classes of EPs

(Ginzburg et al., 2014) propose to distinguish BL OCMs from FL OCMs essentially in terms of distinct issues whose accommodation into QUD they give rise to:

(16) a. BLDs address the issue of *what did A mean with* u_0

b. FLDs address the issue of *what word should A follow* u_0

We can use this idea to offer a basic characterization of EPs compatible with BL OCMs, FL OCMs, or both. By ‘p raising q’ we assume a notion of erotetic entailment (Wiśniewski, 2013):

(17) a. An EP E is BL if content(E) raises the issue *what did A mean with* u_0

b. An EP E is FL if content(E) raises the issue *what word should A follow* u_0

Let us consider two rather clear cases for BL and FL EPs, respectively—‘No/Non’ and ‘I don’t know/Je sais pas’, assuming the following hypothesized contents:

(18) a. ‘No’ $\mapsto_{content}$ I didn’t want to utter u_0.

b. ‘I dont know’ $\mapsto_{content}$ I don’t know what the content of the next utterance should be.

These indeed seem to validate (17). A similar case could be made for ‘Or’ and ‘Voilà’:

(19) a. ‘Or’ $\mapsto_{content}$ There is an alternative to uttering u_0.

b. ‘Voilà’ $\mapsto_{content}$ *That* is what the content of the next utterance should be.

(17) suggests the difficulty in having an EP which is genuinely BL and FL. Empirically these seem to be the fillers whose meaning has typically been explicated in terms of difficulty to make a subsequent utterance (Clark and FoxTree, 2002).

Now it is somewhat facile to engage in content assignations such as (18) and (19). As we have seen, apart from fillers, at least in English and French there seem to be no purpose built EPs. While a realistic grammar will arguably have lexical entries for uses as EPs, these need to be derived or relatable in general ways to their other uses as connectives. We exemplify here two cases, leaving for future work the formulation of a general ‘lexical rule’ or similar.

(Ginzburg et al., 2014), following (Cooper and Ginzburg, 2011), proposed that ‘No’ as an EP is an instance of a bouletic lexical item, exemplified in (20):

(20) a. [A opens freezer to discover smashed beer bottle] A: No! (*I do not want* this *(the beer bottle smashing) to happen*)

b. [Little Billie approaches socket holding nail] Parent: No Billie (*I do not want* this *(Billie putting the nail in the socket) to happen*)

They proposed such a use has the lexical entry in (21):

(21)
$$\begin{bmatrix} \text{PHON} : \texttt{no} \\ \text{CAT.HEAD} = \textit{interjection} : \text{syncat} \\ \text{DGB-PARAMS} = \begin{bmatrix} \text{sit1} : \text{Rec} \\ \text{spkr} : \text{Ind} \end{bmatrix} : \text{RecType} \\ \text{CONT} = \neg\text{Want(spkr,sit1)} : \text{Prop} \end{bmatrix}$$

Its instantiation as an EP can be proposed as (22):

$$(22)\ \left[\begin{array}{ll} \text{PHON} : \texttt{no} & \\ \text{CAT} = \textit{interjection} : \text{syncat} & \\ \text{DGB-PARAMS} : & \left[\begin{array}{l} \text{spkr : IND} \\ \text{addr : IND} \\ \text{MaxPending : LocProp} \\ \text{u0 : LocProp} \\ \text{c1: member(u0,} \\ \text{MaxPending.sit.constits)} \\ \text{rest : address(spkr,addr,} \\ \text{MaxPending)} \end{array}\right] \\ \text{CONT} = \neg\text{Want(spkr,u0) : Prop} & \end{array}\right]$$

By the same token, one could postulate a phrasal description—omitting its phrasal syntactic description—for ‘Je sais pas’: the speaker does not know p, where p? is the currently maximal element of QUD:[2]

$$(23)\ \left[\begin{array}{ll} \text{PHON} : \texttt{je sais pas} & \\ \text{CAT.HEAD} = \textit{verbal} : \text{syncat} & \\ \text{DGB-PARAMS} : & \left[\begin{array}{l} \text{spkr : IND} \\ \text{addr : IND} \\ \text{MaxQUD= p? : Question} \end{array}\right] \\ \text{CONT} = \neg\text{Know(spkr,p) : Prop} & \end{array}\right]$$

Its instantiation as an EP expressing (18b) could be postulated as (24):

$$(24)\ \left[\begin{array}{ll} \text{PHON} : \texttt{je sais pas} & \\ \text{CAT.HEAD} = \textit{verbal} : \text{syncat} & \\ \text{DGB-PARAMS} : & \left[\begin{array}{l} \text{spkr : IND} \\ \text{addr : IND} \\ \text{MaxPending : LocProp} \\ \text{u0 : LocProp} \\ \text{c1: member(u0,} \\ \text{MaxPending.sit.constits)} \\ \text{q =} \\ \lambda x\ \text{MeanNextUtt(spkr,u0,x)} \\ \text{p : Prop} \\ \text{c1 : Resolve(p,q)} \end{array}\right] \\ \text{CONT} = \neg\text{Know(spkr,p) : Prop} & \end{array}\right]$$

6 Conclusions and Future Work

In this paper we have examined the distribution and basic semantic properties of editing phrases in English and French on the basis of the Switchboard and Rhapsodie corpora, respectively. On the one hand, the data we provide demonstrates that the distribution of EPs is highly language–specific. On the other hand, there is evidence that certain semantic properties of EPs are preserved under translation. This provides support for the view that EPs, and more generally, disfluency / Other Communication Management–containing utterances constitute part of the acquired grammatical competence of English and French speakers. This in contrast, for instance, to the distribution of coughs, hiccoughs, and sneezes of speakers of the same languages, which we hypothesize to be roughly similar across distinct languages.

Future work should involve scaling up the corpus description both within and across languages. We plan to develop a far more detailed and systematic account of the relationship between the EP and non-EP use and to implement this in an incremental grammar.

Acknowledgments

Many thanks to three reviewers for SemDial 2015 for their comments. We acknowledge support by the French Investissements d’Avenir–Labex EFL program (ANR-10-LABX-0083) and by the Disfluences, Exclamations, and Laughter in Dialogue (DUEL) project within the *projets franco-allemand en sciences humaines et sociales* of the Agence Nationale de Recherche (ANR) and the Deutsche ForschungGemeinschaft (DFG).

---

[2] We assume some such restriction on p exists for the antecedent of the null propositional object, but this is probably more intricate than simply MaxQUD.

References

Jens Allwood, Elisabeth Ahlsn, Johan Lund, and Johanna Sundqvist. 2005. Multimodality in own communication management. In *Proceedings of the second Nordic WOrkshop on Multimodal COmmunication*, Gothenburg.

Roxane Bertrand, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3):1–30.

Claire Blanche-Benveniste. 1984. La dénomination dans le français parlé: une interprétation pour les répétitions et les hésitations. *Recherches sur le Français Parlé Aix-en-Provence*, (6):109–130.

Herbert Clark and Jean FoxTree. 2002. Using *uh* and *um* in spontaneous speech. *Cognition*, 84(1):73–111.

Robin Cooper and Jonathan Ginzburg. 2011. Negation in dialogue. In *Proceedings of SemDial 2011 (Los Angelogue)*, pages 130–139.

Michel De Fornel and Jean-Marie Marandin. 1996. L'analyse grammaticale des auto-réparations. *Le Gré des Langues*, 10:8–68.

Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong, and Paola Pietrandrea. 2012. Intonosyntactic data structures: the rhapsodie treebank of spoken french. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 85–94. Association for Computational Linguistics.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):1–64.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for conversation*. Oxford University Press, Oxford.

E Matthew Husband. 2015. Self-repairs as right node raising constructions. *Lingua*, 160:20–37.

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, Atanas Tchobanov, et al. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken french. In *Language Resources and Evaluation Conference*.

Willem J. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(4):41–104.

Bertille Pallaud, Stéphane Rauzy, and Philippe Blache. 2013. Auto-interruptions et disfluences en français parlé dans quatre corpus du cid. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, (29).

Elizabeth E. Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California at Berkeley, Berkeley, CA.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Jean E Fox Tree and Josef C Schrock. 2002. Basic meanings of you know and i mean. *Journal of Pragmatics*, 34(6):727–747.

Andrzej Wiśniewski. 2013. *Questions, Inferences, and Scenarios*. College Publications, London, England.

Poster Presentations

The significance of silence:
Long gaps attenuate the preference for ‘yes’ responses in conversation

Sara Bögels[1], Kobin H. Kendrick[1], and Stephen C. Levinson[1,2]
[1]MPI for Psycholinguistics, P.O. Box 310, 6500 AH, Nijmegen, The Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
`{sara.bogels, kobin.kendrick,stephen.levinson}@mpi.nl`

Abstract

In conversation, negative responses to invitations, requests, offers and the like more often occur with a delay – conversation analysts talk of them as dispreferred. Here we examine the contrastive cognitive load ‘yes’ and ‘no’ responses make, either when given relatively fast (300 ms) or delayed (1000 ms). Participants heard mini-dialogues, with turns extracted from a spoken corpus, while having their EEG recorded. We find that a fast ‘no’ evokes an N400-effect relative to a fast ‘yes’, however this contrast is not present for delayed responses. This shows that an immediate response is expected to be positive – but this expectation disappears as the response time lengthens because now in ordinary conversation the probability of a ‘no’ has increased. Additionally, however, 'No' responses elicit a late frontal positivity both when they are fast and when they are delayed. Thus, regardless of the latency of response, a ‘no’ response is associated with a late positivity, since a negative response is always dispreferred and may require an account. Together these results show that negative responses to social actions exact a higher cognitive load, but especially when least expected, as an immediate response.

1 Introduction

Most natural language use occurs in conversational contexts, in which paired initiating and responding actions are prevalent (e.g., request-granting, greeting-greeting; Schegloff, 2007). Responses after an initiating action are rarely equal: the initiating action is constructed to expect a particular response, which is usually positive, such as granting a request (Pomerantz & Heritage, 2013). This unmarked response tends to have a simple form (e.g., *yeah* or *sure*). In contrast, negative responses which reject or decline the initiating action tend to be delayed in time and to occur with prefaces and accounts (e.g., *Well, no, I’m too tired*) for the negative response (Pomerantz & Heritage, 2013). This structural and temporal asymmetry has been called ‘preference organization’ (Levinson, 1983); for many action pairs positive responses are 'preferred' and negative responses 'dispreferred'. It was first noted in qualitative research (e.g., Heritage, 1984) that preferred responses generally occur quickly after the initiating action, whereas dispreferred responses are more often delayed. Kendrick and Torreira (2015) quantified these observations in a corpus study on English. They found that preferred responses occurred most prevalently, but after around 700 ms, dispreferred responses became more frequent.

Language comprehension research often uses event-related brain potentials (ERPs); EEG responses to specific events. ERP studies on language comprehension have made extensive use of the N400 component, the amplitude of which has been found to vary with the expectation of a word with respect to its preceding context (e.g., Kutas & Hillyard, 1980); the more expected a word, the smaller the N400. In the present study, we looked at ERP responses to fast and delayed preferred and dispreferred responses. We hypothesized that a preferred response ('yes') should be more expected than a dispreferred response ('no'), probably leading to a larger N400 for 'no', especially after a short gap. However, after a long gap a dispreferred response (‘no’) should be less exceptional, leading to a smaller N400 effect or even a reversal in that case.

2 Methods

Thirty-two participants (8 males) entered the analyses (mean age: 21.8). We took 120 requests, offers, invitations, and proposals from recorded telephone calls in the Corpus of Spoken Dutch (CGN, Oostdijk, 2000) and cross-spliced them with either a 300 ms or a 1000 ms gap of recording noise, followed by either a 'yes' or a 'no' response from elsewhere in the corpus. Conditions were counterbalanced such that each participant heard each initiating action only once. Participants first read a context sentence, followed by the two-turn sequence. This was followed by a comprehension question in 20% of trials. EEG was recorded from 61 active Ag/AgCI electrodes using an actiCap (e.g., Magyari et al., 2014). Trials with blinks identified by eye electrodes were discarded before analysis, using Fieldtrip (Oostenveld et al., 2011). A cluster-based approach was used for statistical analysis (Maris & Oostenveld, 2007).

3 Results and Discussion

Figure 1 shows the ERPs for a representative electrode (Cz), time-locked to response onset for the 4 conditions. An interaction between gap length and response type in the N400 window (300-500 ms, p = .032) was resolved to show an N400 effect for 'no' versus 'yes' after a short gap (p = .006), but no difference after a long gap. Listeners apparently change their expectations about a preferred vs. dispreferred response purely based on the length of silence between two turns. In particular, listeners expect a preferred response ('yes') rather than a dispreferred one ('no') after a short gap, but these expectations converge after a long gap. Our finding that a dispreferred response does not become *more* expected than a preferred response after a long gap might be related to the normativity of preferred responses: they favor the accomplishment of the activity and have been associated with social solidarity and social affiliation (Heritage, 1984). A general bias might exist towards socially affiliative responses, which might balance the delay-induced expectation of a dispreferred response, leading to a net outcome of no N400 effect.

No interaction between gap length and response type was found after 500 ms. Instead, a main effect of response type showed a larger frontal late positivity for 'no' compared to 'yes' responses, irrespective of gap length (p = .002). As mentioned, the preference system biases expectations towards socially affiliative responses, so that there are extra social consequences for rejections. For that reason, rejections are often accompanied by accounts and explanations (e.g., *No, I can’t, I have to work*). The rejections in our stimuli, in contrast, had no such accounts attached because we had to control our stimuli and match them in length to the acceptances and compliances. The flat 'no' responses in our experiment might thus be seen as rude – leading to stronger social and cognitive consequences (see also e.g., Leuthold et al., 2015).

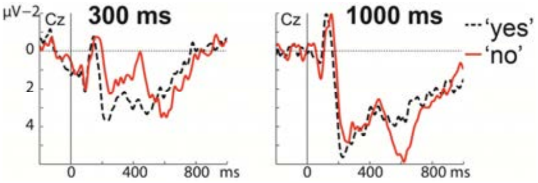


Figure 1. ERPs for 'yes' and 'no' responses after a 300 (left) and 1000 ms gap (right).

Acknowledgments

This work was supported by an ERC Advanced Grant (269484 INTERACT) to Stephen C. Levinson.

References

Heritage, J. (1984). A change-of state token and aspects of its sequential placement. *Structure of social action: Studies in conversation analysis*, 299-345.

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203-205.

Leuthold, H., Kunkel, A., Mackenzie, I. G., & Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. *Social cognitive and affective neuroscience*, nsu151.

Levinson, S. C. (1983). *Pragmatics* (Cambridge textbooks in linguistics).

Magyari, L., Bastiaansen, M. C., de Ruiter, J. P., & Levinson, S. C. (2014). Early Anticipation Lies behind the Speed of Response in Conversation. *Journal of Cognitive Neuroscience, 26*, 2530-2539.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods, 164*(1), 177-190.

Oostdijk, N. (2000). *Het Corpus Gesproken Nederlands.*

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience, 2011*, 1.

Pomerantz, A., & Heritage, J. (2013). Preference. *The handbook of conversation analysis*, 210-228.

Schegloff, E. A. (2007). *Sequence organization in interaction: Volume 1: A primer in conversation analysis* (Vol. 1): Cambridge University Press

Within reason: Categorising enthymematic reasoning in the balloon task

Ellen Breitholtz and Christine Howes
Centre for Language Technology
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
`ellen.breitholtz@ling.gu.se`
`christine.howes@gu.se`

Abstract

In dialogue, people often use reasoning that relies on information not explicitly present in the discourse, or *enthymemes*. We report on a preliminary corpus study to categorise the enthymematic arguments used in text chat discussions of a moral dilemma; the *balloon task*.

1 Introduction

When engaging in conversation we sometimes use arguments. This tendency is stronger in some types of dialogue, but is present even in everyday conversation, as discussed in (Breitholtz, 2014) and (**?**). The type of arguments we use in conversation are almost always *enthymematic*, that is they need to be supplied with information that is not explicitly present in the discourse, but only in the minds of the langauge users. In rhetoric it is thus important to choose the enthymemes you use as a speaker to tap into patterns of reasoning that are recognised and accepted by the audience. In this paper we try to investigate the types of arguments used in 11 argumentative dialogues on the same topic.

2 Method

2.1 Corpus

The corpus of data consisted of 11 online, text-based dialogues between pairs of native English speakers, collected using the DiET chat tool (Healey et al., 2003), at Queen Mary University of London.

Participants discussed the *balloon task* – an ethical dilemma requiring agreement on which of three passengers should be thrown out of a hot air balloon that will crash, killing all the passengers, if one is not sacrificed. The choice is between Nick, a scientist, who believes he is on the brink of discovering a cure for cancer, Susie, a woman who is 7 months pregnant, and Tom, her husband, the pilot. This task has been used for studying many aspects of dialogue, as it is known to stimulate discussion (Howes et al., 2011).

2.2 Annotations

The corpus was annotated for arguments regarding who to save and who to throw out. For each claim that someone should be saved and thrown out we also noted what seemed to be at the core of the enthymeme, that is, the gist of the argument. For example, one participant wants to throw Susie out with the motivation that the potential of her unborn baby is uncertain.

3 Results and Discussion

		Nick	Susie	Tom	Total
All	Throw	184	27	51	262
	Save	75	78	78	231
Threw Nick	Throw	132	14	22	168
	Save	39	58	57	154
Threw Tom	Throw	18	6	25	49
	Save	14	9	1	37

Table 1: Number of turns containing a reason to save or throw each person

Of 1983 turns in total, 1496 (26%) contained reasoning about who to keep in the balloon or who to throw out. Interestingly, as can be seen from Table 1, although participants supply approximately as many turns containing arguments to save each person, they provide a far higher proportion of turns which offer arguments for throwing Nick. This is in line with the fact that of the 11 pairs, 7 opted to throw Nick. 3 pairs opted to throw Tom, and one pair did not reach agreement with one participant opting to throw Nick, and the other Susie.

3.1 Arguments used

As shown in Table 2, there are a number of different arguments employed by participants in justifying their decision of who to throw out of the balloon and who to save. Some of these occur in most of the dialogues, such as the reasoning that Nick only believes he is on the brink of a cure for cancer (see e.g. 1), whilst others are rarer, such as the reasoning that the balloon losing height is Tom's fault (see e.g. 2). Similarly, some reasons are specifically tailored to one of the people, and others can be used to justify different conclusions, such as the speculation about who is heaviest 'weight' (see the examples in 3, taken from different dialogues).

	Nick	Tom	Disagree	Total
Nick can save lives	30	12	21	63
Tom can fly	41	13	5	59
Nick only believes	29	10	19	58
Nick has notes	28	1	2	31
Susie family	25	1	5	31
Nick has team	22	4	4	30
Susie is two people	17	8	5	30
Tom family	18	9	3	30
Tom can explain flying	8	10	0	18
Emotive	12	2	2	16
Least important	10	1	3	14
Nick is nice	12	1	0	13
Susie is teacher	9	0	4	13
Tom's fault	6	6	0	12
Nick family	7	1	1	9
Nick isn't nice	9	0	0	9
Unborn baby potential	5	0	4	9
Weight	8	2	0	10
Media response	8	0	0	8
Nick could fly	3	0	3	6
Susie is weak	6	0	0	6
Nick might be father	1	2	0	3
Susie Tom couple	1	2	0	3
Nick can explain	0	0	2	2
Nick can help after	2	0	0	2
Nick is old	2	0	0	2
Tom duty	0	1	0	1
Total	319	86	83	488

Table 2: Gloss of reasons given

(1) A he 'believes' he is on the brink
A his research might be dudd
R he could be bluffing

(2) F cos hes a balloon pilot and therefore he would of known the consequences of the balloon in the first place

(3) T if the dr is twice the size of tom, that would guarantee their extra height
P but then the woman may also be heavier because she is carrying a child.....

4 Conclusions and Future work

We intend to investigate whether this categorisation of enthymematic reasoning in the balloon task is robust, and can be used to predict or influence who participants will throw off the balloon. We will extend the preliminary work presented here to face-to-face dialogue, and also to see if the reasoning deficits described in patients in schizophrenia (Langdon et al., 2010) can be accounted for by enthymemes and their underlying topoi. We propose to test this using an existing corpus, using a variant of the balloon task between either three healthy control participants or two healthy control participants and one patient with schizophrenia (Lavelle et al., 2012).

Acknowledgments

Thanks to Shauna Colcannon for the availability of the corpus.

References

Ellen Breitholtz. 2014. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis, University of Gothenburg, September.

P. G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, Massachusetts, August.

Christine Howes, Matthew Purver, Patrick G. T. Healey, Gregory J. Mills, and Eleni Gregoromichelaki. 2011. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311.

Robyn Langdon, Philip B Ward, and Max Coltheart. 2010. Reasoning anomalies associated with delusions in schizophrenia. *Schizophrenia bulletin*, 36(2):321–330.

Mary Lavelle, Patrick G. T. Healey, and Rose McCabe. 2012. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia Bulletin*.

Context influence vs efficiency in establishing conventions: Communities do it better

Lucia Castillo
University of Edinburgh
s1354576@sms.ed.ac.uk

Holly Branigan
University of Edinburgh
holly.branigan@ed.ac.uk

Kenny Smith
University of Edinburgh
kenny.smith@ed.ac.uk

Abstract

The emergence of communicative conventions in human groups is believed to be governed by both local forces of salience and precedence, and global forces pointing to a convergence onto the most frequently encountered alternative (Garrod and Doherty, 1994). In the present study we tried to answer two questions: 1) what is the influence of context over the establishment of conventions? And 2) are communities as sensitive to that influence as pairs? Using a maze game task, we compared communities and pairs of participants in two different contexts: a regular context, where the maze layout is closer to a grid, and an irregular context, where the layout resembles an irregular shape. We predicted that regular layouts would cue the use of more abstract description schemes to refer to locations in the maze, while irregular layouts would cue the use of more concrete schemes. Our results show that participants in the irregular context were more likely to use concrete description schemes in the first game in both pairs and communities, but while pairs of participants maintained this choice over the following games, communities moved towards the more efficient abstract description schemes. These results show that the influence of context can be overcome by communities, and that the most frequently encountered initial scheme is not necessarily kept if there are more efficient alternatives available.

1 Introduction

We investigated the effect of the context in which communication takes place on the nature of the emerging conventions, comparing pairs versus communities of players in a maze game task. Pairs of participants, communicating over a chat interface (Healey and Mills, 2006; Mills, 2014), had to jointly identify and locate tangram figures distributed in a maze. Both participants in each pair had the same maze structure but the figures were placed in different positions. The task forced them to describe and agree on the positions of the tangrams. We tested whether differences in the regularity of the maze would prompt participants to use different description schemes to refer to locations in the maze. We predicted that more regular (i.e. grid-like) maze would cue the use of abstract description schemes which make use of the grid-like appearance of the maze (i.e. by referring to positions in terms of row and column numbers), while more irregular mazes (characterized by e.g. irregular protrusions) would cue the use of more concrete description schemes relying on salient features of the mazes as reference points. Moreover, we predicted that, as abstract schemes are more efficient in both regular and irregular contexts, communities would tend to move towards the use of abstract schemes, while pairs would be bounded by salience and precedence to the schemes they used in the first rounds.

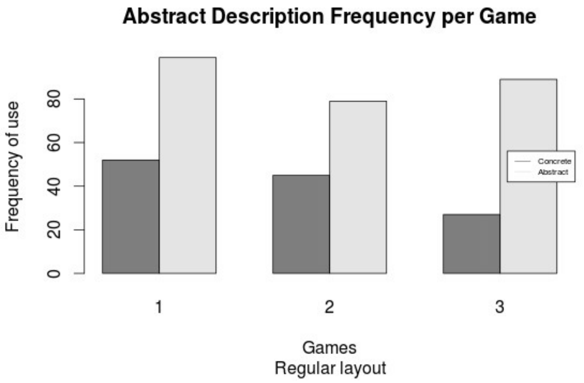
2 Methods and procedure

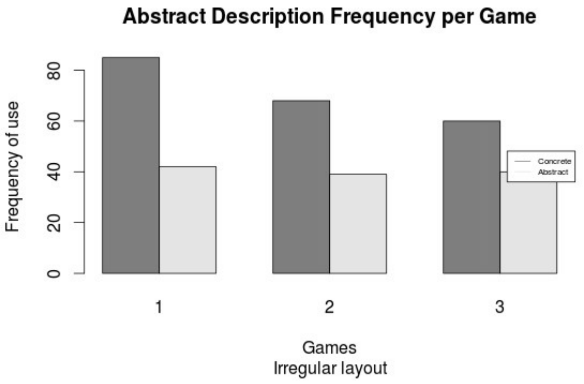
14 pairs and 8 four-people groups played over a chat interface for 3 games each. In the pairs setting, pairs of participants played together for 3 games, while in the communities setting, participants played with a different member of the groups in each of the 3 games, forming an emergent community. Each maze was based on a 7x7 grid and contained a similar number of squares. We developed a measure of maze regularity, based on mean square density, to select two samples of regular and irregular mazes. Pairs/groups would play on either a regular or an irregular maze for 3 games. On each maze, players had to identify and describe the position of 6 tangram figures. The figures were the same

for all pairs and groups. Once both players had identified and selected the position of a chosen figure in each other's mazes, the figure disappeared, and they moved to the next figure until all figures were gone.

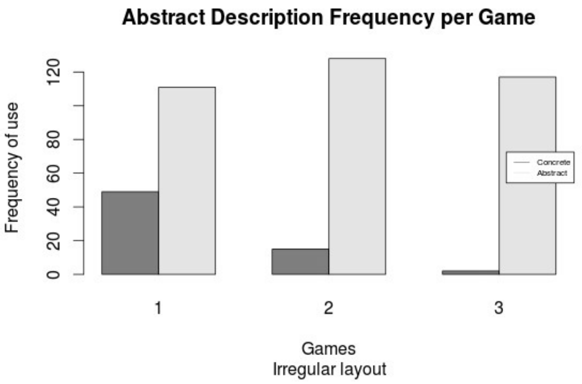
3 Results

Regular mazes resulted in the use of more abstract description schemes, and irregular mazes were associated with concrete description schemes: in the first game, the probability of using an abstract description was significantly higher for pairs in a regular layout than for pairs in an irregular layout, across both pairs and communities conditions (pairs were 0.5 times more likely to use an abstract description in a regular layout, compared to irregular, in the pairs condition, and 0.77 times in the communities condition). However, pairs showed a similar difference in games 2 and 3, which shows that they maintained their description schemes across games.





Communities, on the other hand, showed the predicted behaviour, with groups in different layouts using different description schemes over the first game, but converging over the abstract description schemes over the second and third games. The difference between conditions was significant for the first game (regular vs irregular layout, use of abstract description schemes, but became neutralised through the games, with no significant difference between conditions by game 3.



4 Discussion

These results suggest that context affects participants' choice of reference scheme, with regular contexts cueing the use of abstract schemes, and concrete, maze-specific schemes being preferred in irregular contexts; but that this selection is only maintained in pair-wise settings. Communities, on the other hand, move away from concrete schemes –even in the more salient irregular layout– towards abstract, more efficient schemes, as participants interact with different partners. This increased efficiency in communities shows how a 'better' alternative can become established as convention even when it was not the most salient option in a given context.

References

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. Cognition, 53, 181±215.

Healey, P. G. T., & Mills, G. J. (2006). Participation, precedence and co-ordination in dialogue. In Proceedings of the 28th annual conference of the Cognitive Science Society. Vancouver, Canada.

Mills, G. J. (2014). Dialogue in joint activity: Complementarity, convergence and conventionalization. New Ideas in Psychology, 32, 158-173.

Are response particles not well understood? Yes/No, they aren't! An experimental study on German *ja* and *nein*

Berry Claus, Marlijn Meijer, Sophie Repp, and Manfred Krifka
Humboldt-Universität zu Berlin
`{berry.claus|meijeram|sophie.repp|manfred.krifka}@hu-berlin.de`

Abstract

Response particles such as English *yes* and *no* are frequently used in dialogues, to respond to questions or assertions. However, while the use of *yes* and *no* is straightforward in responses to non-negated antecedent clauses, it is not clear-cut with negated antecedent clauses. For example, to agree with an assertion such as *Jim doesn't snore*, both *yes* and *no* can be used (*Yes/No, he doesn't*). The same holds for the German response particles *ja* and *nein* (roughly corresponding to *yes* and *no*). In the present study, we investigated preference patterns for German *ja* and *nein* as responses to negated assertions. Our results revealed two distinct subgroups of participants. One subgroup, approx. 70% of the participants, showed a preference for *ja* over *nein* as agreeing responses to negated antecedents, whereas the other subgroup, approx. 30% of the participants, showed a preference for *nein* over *ja*. To account for this finding, we put forward an ellipsis analysis and propose that the two groups differ with respect to the meaning of *nein*.

1 Introduction

Dialogues are rife with response particles such as *yes* and *no*, which are a short means of answering yes/no questions or expressing agreement/disagreement with assertions. However, their use and interpretation is clear-cut only for non-negated antecedent clauses, such as *Jim snores*. Here, *yes* and *no* are used complimentarily. For negated antecedents, such as *Jim doesn't snore*, *yes* and *no* are not complimentary. They can both be used in disagreeing responses to negated antecedents (*Yes/No, he does*) and they can both be used in agreeing responses (*Yes/No, he doesn't*). The latter also holds for the German response particles *ja* and *nein* (roughly corresponding to *yes* and *no*). The German response particle system differs from English in that it is a three particle system. Besides *ja* and *nein*, it includes the specialized particle *doch*. *Doch* is a dedicated particle for disagreeing responses to negated antecedents, whereas for agreeing responses, both, *ja* and *nein*, can be used (see 1).

(1) A: *Jim schnarcht nicht*. ('Jim doesn't snore)
B: i. *Ja.* (= He doesn't snore.)
ii. *Nein.* (= He doesn't snore.)
iii. *Doch.* (= He does snore.)

Two recent approaches to response particles, the semantic-syntactic feature model of Roelofsen & Farkas (to appear; =R&F) and the anaphor account of Krifka (2013) allow for predictions as to preference patterns for *ja* and *nein* as responses to negated antecedents. In a nutshell, R&F propose for disagreeing responses that both *ja* and *nein* are blocked due to the presence of *doch* in the system. In contrast, Krifka supposes that *doch* blocks *ja*, whereas *nein* is not blocked, albeit dispreferred. For agreeing responses, R&F predict a general preference for *nein* over *ja*, and Krifka predicts a preference for *nein* over *ja* in default contexts.[1]

2 Experimental study

In three acceptability-judgment experiments on responses to negated assertions, participants were presented with short dialogues, as illustrated in Table 1. Each dialogue was preceded by a scene-setting passage, which introduced the two interlocutors and served as the dialogue's context, specifying what the two interlocutors were talking about.[2] The participants' task was to judge the naturalness and suitability of the response in the given dialogue and context on a scale ranging from 1 (very bad) to 7 (very good).

---

[1] Krifka's account implies, contra R&F, that the preference for *ja* or *nein* should be sensitive to the wider discourse context. For contexts, in which the negated proposition expressed by the antecedent is salient (rather than its positive counterpart which is assumed to be the salient proposition by default), Krifka predicts a preference for *ja* over *nein*.

[2] The context was varied to manipulate the saliency of the negated antecedent proposition vs. its positive complement. However, as the data did not show any significant interaction effects involving the factor CONTEXT, only results of analyses obtained from data pooled over the two context conditions are presented here for simplicity.

(Setting: Ludwig and Hildegard have their large garden redesigned)
L: *The gardener hasn't sown the lawn yet.*
H: *Yes/No, he has sown the lawn already./*
Yes/No, he hasn't sown the lawn yet.

Table 1: Sample of the dialogues employed in Expt. 1, translated from German

In **Expt. 1**, we manipulated the factors RESPONSE PARTICLE (*ja*/*nein*), and RESPONSE CLAUSE POLARITY (positive/negative). In the positive response clause conditions, i.e. disagreeing responses, ratings for *ja* were quite low (*M*=2.11) and significantly differed from the ratings for *nein* (*M*=5.34), suggesting, that only *ja* but not *nein* is blocked by *doch*. The results of **Expt. 2**, which included *doch* as an additional level of RESPONSE PARTICLE demonstrated significantly higher ratings for *doch* (*M*=6.76) compared with *nein* (*M*=3.84) and *ja* (*M*=1.81), and replicated the significant difference between *nein* and *ja*. In the negative response clause conditions of **Expt. 1**, i.e. agreeing responses, *ja* (*M*=6.09) was rated significantly higher than *nein* (*M*=4.80). This pattern was replicated in **Expt. 3**, where the responses did not include a follow-up phrase, but were bare particles.[3] As in Expt. 1, *ja* (*M*=5.91) received significant higher ratings than *nein* (*M*=4.24). Thus, contra both R&F's feature model and Krifka's anaphor account, the results of Expt. 1 and 3 indicate a general preference for *ja* over *nein* as agreeing responses to negated antecedents rather than for *nein* over *ja*. However, a closer inspection of the data revealed two distinct subgroups of participants. About 70% of the participants of Expt. 1 and 3 showed the unpredicted pattern of higher ratings for *ja* than for *nein* ("Yes-group"). In contrast, about 30% of the participants in both experiments, showed the reverse pattern, i.e. higher ratings for *nein* compared to *ja* ("No-group").

3 An ellipsis account

To account for the observed data pattern, we propose an ellipsis account. Syntactically, we analyse *ja, nein* and *doch* as operators that operate on the TP, which is a copy of the antecedent, and is obligatorily elided. With respect to the opposite preference patterns for the two subgroups, we suggest that the two groups apply different response systems: truth-value vs. polarity based (Jones, 1999). The "Yes-group" uses a truth-value based system with *ja* signalling the truth (and *nein* the falsity) of the antecedent, whereas the "No-group" uses a polarity based system with *nein* signalling a negative response polarity (and *ja* a positive one). Formally, this difference can be modelled in a parsimonious way by assuming that the two groups differ only in the meaning of *nein* (see Table 2).

Both groups	⟦*ja*⟧ = λp.p ⟦*doch*⟧ = λp:"p is negative".¬p
"Yes-group" ⟦*nein*⟧ = λp.¬p	**"No-group"** ⟦*nein*$_1$⟧ = λp.¬p ⟦*nein*$_2$⟧ = λp:"p is negative". p
Note: p = antecedent proposition; *doch* and *nein*$_2$ have the presupposition that p is negative	

Table 2: Proposed meanings for *ja, nein*, and *doch*

As a brief illustration of the proposed semantics consider the decisive case of agreeing responses to negated antecedents (e.g., A: *John doesn't snore*. Intended response of B: He doesn't snore). For the "Yes-group", *ja* is the only response particle that expresses the intended meaning (=antecedent proposition). For the "No-group", both *ja* and *nein*$_2$ express the intended meaning, with *nein*$_2$ being preferred over *ja* due to *Maximize presupposition* (Heim, 1991).

To conclude: our experimental study revealed two subgroups of participants, differing in the preference pattern for *ja* and *nein* as agreeing responses to negated assertions. As a preliminary proposal, we put forward an ellipsis account, deserving further study.

Acknowledgments
This research is supported by DFG in the priority program Xprag.de.

References

Heim, I. (1991). Artikel und Definitheit. In A. von Stechow & D. Wunderlich (eds.), *Semantik – Semantics* (pp. 487-535). Berlin: de Gruyter.

Jones, B. M. (1999). *The Welsh answering system*. Berlin: Mouton de Gruyter.

Krifka, M. (2013). Response particles as propositional anaphors. *Proceedings of SALT* 23, 1-18.

Roelofsen, F. & Farkas, D. (to appear). Polarity particle responses as a window onto the interpretation of questions and assertions. *Language*.

---

[3] To make clear whether a bare *ja* or *nein* should be taken as an agreeing response, the scene-setting passages in Expt. 3 contained information on the 'epistemological state' of the responding person (e.g. *The gardener told Hildegard that he would sow the lawn in a couple of days*).

KILLE: Learning Objects and Spatial Relations with Kinect

Erik de Graaf
University of Gothenburg
gusdegrer@student.gu.se

Simon Dobnik
University of Gothenburg
simon.dobnik@gu.se

Abstract

We present a situated dialogue system designed to learn objects and spatial relations from relatively few examples, based on camera imagery and dialogue interaction with a human partner. We also report on the baseline evaluation of the system.

1 Introduction

Grounding, the linking of real world objects and situations involving objects to their computational semantic representations, is a necessary step for meaningful interaction with robots (Roy, 2005). Systems that operate within the real world will often encounter novel situations and word usages and therefore they will need to *learn* new semantic representations. In contrast to state of the art systems that work with large databases of images to learn from, our system tries to learn grounded meanings of objects and spatial relations from a very few examples presented to the system in situated interactive learning. Our long term goal is to investigate how various dialogue interaction strategies with a human can leverage the sparsity of observable data.

2 Object and scene recognition

The hardware used is a Kinect 3d camera, connected to a computer. The camera is mounted stationary to a table on and over which objects are presented to the system. The Freenect drivers[1] are used to capture data from the camera and to forward them to the Robot Operating System (ROS) framework (Quigley et al., 2009). The dialogue is managed by OpenDial (Lison, 2014), including speech recognition and speech synthesis. Rules for the dialogue system are written in OpenDial's own XML format. Objects are learned by storing the recognized SIFT features or SIFT descriptors (Lowe, 2004) of each object instance that are calculated from the frames the camera forwards. Before learning and recognizing objects the background is removed. This way we remove distracting features not belonging to the object in focus. SIFT-features are well known and frequently used in object recognition, for their rotation- and scale-invariance and performance in matching to other sets of features. The SIFT descriptors are represented as multi-dimensional vectors, abstracted from important points in an image, such as corners or edges. Once objects have been learned new objects are classified by finding the category of the most closely matching object in terms of SIFT. Objects are matched by finding the highest harmonic mean of two measures. In the first measure the number of visual features matched between the recognized and a learned object is divided by the number of features of the learned object, whereas in the second it is divided by the number of features of the recognized object. The category of the stored object with the highest score is picked as the name of the object recognized. For spatial relations the locations of objects are represented as average x, y and z coordinates of detected SIFT features.

3 Conversational strategies

The system learns objects either by being presented with them and told what they are (e.g. *This is a cup*) or by receiving feedback on an utterance it just made (*That's correct*). When the system hears a question such as *What is this?* (or a variation on this) it responds by also describing the certainty of its belief (*The object is thought to be a book, but it might also be a mug*). It can learn spatial relations when it recognizes both of the objects mentioned (*The book is on the right of the mug*). The system is also able to learn from feedback, confirmations of a human partner whether something was correct or not. The system may occasionally mishear the name of an object. The name can be unlearned right after learning (by saying *That is not what I said*), unlearned later (*Forget cup*) or re-learned to attach a new name to the previously learned object (*I said a book*). The system will occasionally ask the user for more examples of an object or spatial relation that it has too little knowledge of, but assumes the tutor takes the lead

---

[1] `http://openkinect.org/wiki/Main_Page`

	Accuracy	Accuracy cumulative
Round 1	96%	96%
Round 2	94%	95%
Round 3	96%	95.3%
Round 4	98%	96%

Table 1: Accuracy of recognition after the different testing rounds.

again right after that. This happens at random after a response or acknowledgement from the system.

4 Baseline evaluation

In the current experiment we test object recognition without human feedback. This will serve as a baseline for our forthcoming work where we will be testing incrementally more sophisticated interaction strategies that were described in the previous section. Ten objects are shown to the system for four rounds. After each presentation the system is queried for that object category. Note that although the object has not moved the system will make the classification from a new sensory scan. At each round the objects are placed in the same order and with approximately the same position and orientation.

5 Results and discussion

The accuracy of object recognition at each round as well as the cumulative accuracy over several rounds is presented in Table 1. These results show that accuracy of the system is very high and that it improves when more instances are learned. Table 2 shows the object matching scores over all object matches. The first column indicates objects presented to the system. The second column shows the average maximal matching scores (AMMS) with an object from the correct category (which may not be the winning one) over the four rounds, and the third column shows the corresponding standard deviations. High scores tell us that objects are easy recognisable, whereas low scores indicate that their recognition is more difficult. The fourth column shows the average overall matching scores (AOMS) against all object models, and the last column shows their standard deviations. This column demonstrates how much an object looks like any other object. Ideally, as we want objects to be uniquely distinguishable, AMMS should be high, while AOMS should be low.

Object	AMMS	Std. dev.	AOMS	Std. dev
Apple	.34	.07	.12	.10
Banana	.36	.07	.12	.10
Bear	.26	.06	.11	.06
Book	.50	.07	.19	.12
Cap	.15	.06	.10	.05
Car	.41	.06	.13	.11
Cup	.33	.10	.11	.09
Paint can	.22	.04	.11	.05
Shoe	.32	.01	.11	.08
Shoebox	.38	.07	.22	.11

Table 2: Object score and standard deviation.

6 Future work

In the immediate future we will examine the effects of varying object orientation and switching objects for other objects of the same category on the rate of learning. We will also test the learning of spatial relations. A change of interaction strategy will also be examined, starting with the contributions of feedback on learning and recognizing. An object ontology could also be implemented. The system could actively query users to gain information about how general the used term is, whether it is the name of a category or an object. As the learned databases are exportable, users could exchange these databases to increase the number of objects and spatial relations a system can recognize. Such a database could be made available on the internet, and divided into categories, depending on where the robot needs to work and what objects it will encounter. As the scale increases, however, it might become feasible to implement recognition with deep convolutional neural networks in favour of SIFT feature detection.

References

P. Lison. 2014. *Structured Probabilistic Modelling for Dialogue Management*. Ph.D. thesis, Univ. of Oslo.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1):170–205.

Demonstrating the Dialogue System of the Intelligent Coaching Space

Iwan de Kok[1,2], **Julian Hough**[1,3], **Felix Hülsmann**[1,4], **Thomas Waltemate**[1,4],
Mario Botsch[1,4], **David Schlangen**[1,3], and **Stefan Kopp**[1,2]
[1]CITEC, [2]Social Cognitive Systems, [3]Dialogue Systems Group,
[4]Computer Graphics & Geometry Processing Group
Bielefeld University
idekok@techfak.uni-bielefeld.de

Abstract

This demonstration presents the current state of the dialogue processing of the Intelligent Coaching Space. This is a multimodal virtual environment in which users are coached in the acquisition of a physical skill. The demonstration highlights the closed interaction loop between the physical action of the user and the responses of the virtual coach.

1 The Intelligent Coaching Space

This demonstration presents the current state of the dialogue processing in the ICSPACE (Intelligent Coaching Space) project at the Cluster of Excellence 'Cognitive Interaction Technology' at Bielefeld University. In this project we are building an immersive, multimodal virtual environment in which users are coached in the process of motor skill acquisition.

A virtual coach observes the user attempting to acquire a motor skill (in our first scenario we focus on squats) and gives incremental instructions and feedback as a human coach would. The domain of physical skill acquisition creates challenges for our dialogue system not present in more traditional pedagogic domains such as tutorial systems. These challenges include fully multimodal input and generation of actions, additional grounding. See (Hough et al., 2015) for a more detailed discussion.

2 Physical Setup

The lab setup of our Intelligent Coaching Space is realized in a Cave Automatic Virtual Environment (CAVE), an immersive 3D Virtual Reality environment with front and floor projection. Users enter our environment wearing 3D glasses and a motion capture suit. They are then tracked by 10 OptiTrack motion capture cameras.

The 3D glasses are tracked to adjust the perspective in our highly responsive custom-built renderer visualizing the scene. In the scene the users see a virtual reflection of themselves in a Virtual Mirror, which is rendered using data from the motion capture suit. Next to the mirror a virtual coach is present that observes the user and instructs the user on how to do a squat.

The Motion Analyzer identifies squats in the stream of motion tracking data and classifies errors made. Based on this information the dialogue system determines the next coaching action.

3 Dialogue Processing

The dialogue system consists of several components which will be discussed in detail below.

The Coaching Strategy Manager is responsible for selecting the next coaching action. It is implemented as a finite state machine making decisions based on an information state. This information state is updated by processing the incoming user input, in this case output of the Motion Analyzer, and also feedback from the Realizer, which informs the Coaching Strategy Manager on the status of its own behaviour.

The information state keeps track of the how many squats have been performed by the user in the current interaction, the errors made during each squat and which phase of the squat the user is currently in.[1] Central to this information state model is the variable *Skills-Under-Discussion*, which unlike traditional Questions-Under-Discussion (QUD) components is not a stack of proposition-based questions, but one of action representations. Based on the current state, the Coaching Strategy Manager selects the most appropriate coaching act, which could be an instruction, demonstration, explanation or feedback.

---

[1] The squat is separated into a preparation phase (assuming the starting position), stroke (going down), strokehold (in the lowest position) and retraction phase (coming up).

Based on the number and severity of errors in the last squat the Coaching Strategy Manager decides whether to address existing errors by pushing the sub-action(s) performed erroneously onto the Skills-Under-Discussion (SkUD). This will modify the coaching acts that can be selected so they specifically target this aspect of the action. Like Ginzburg (2012)'s QUD, SkUD pops its top element in a stack-like fashion once the error has been corrected, and the overall interaction follows a coaching cycle as described in (de Kok et al., 2014).

Based on the decision made by the Coaching Strategy Manager, the Action Pattern Manager activates the Action Patterns required to realize the action of the coach. These Action Patterns are designed to be dynamically created, activated and/or stopped. Currently each action the Coaching Strategy Manager can choose is implemented as its own Action Pattern. All Action Patterns are their own decision makers that are free to produce behaviour fitting the constraints from earlier decision makers, typically the Coaching Strategy Manager.

Each Action Pattern can create its own information flow links to all other parts of our system. For instance, the `Incremental Instruction` pattern directly listens to the output of the Motion Analyzer, bypassing the Coaching Strategy Manager. Note that it can still be deactivated by the Coaching Strategy Manager if it decides on another action. The pattern will produce instructions to improve the on-going squat. Instruction selection is based on the errors detected in the squat and can be restricted by the Coaching Strategy Manager. E.g., if the Coaching Strategy Manager has decided that the maximal Skill-Under-Discussion is X and both an error in X and Y are observed, only an instruction to correct X will be vocalized. If no restrictions are placed, the action pattern is free to make this decision itself. It will continue giving these instructions until either no squat is currently being performed or the Action Pattern is deactivated by the Action Pattern Manager in response to a decision by the Coaching Strategy Manager. When de-activated it will immediately interrupt all its current and planned behaviours.

Other Action Patterns may be more straightforward, simply converting the action selected by the Coaching Strategy Manager into behaviour directly, without listening to any input other than that from the Action Pattern Manager.

Actions produced by the Action Patterns are realized by the AsapRealizer (van Welbergen et al., 2014). It transforms the actions into joint rotations, blend shapes and sound (using CereVoice TTS) which are passed on to the renderer.

4 Demonstration Overview

The demonstration will feature a portable version of the system presented in Section 2, which will highlight some of the challenges in dialogue management presented in Section 1.

To scale down the demonstration, the 3D CAVE environment is reduced to a single monitor. The screen will show the Virtual Mirror on which a virtual reflection of a coachee is displayed. Our virtual coach will stand next to mirror, interacting with the virtual coachee's reflection.

Instead of motion capturing people performing a squat, the demonstration will play prerecorded squats from file. These will be processed by the Motion Analyzer and played back on the screen in the Virtual Mirror. Our virtual coach will incrementally instruct the coachee during playback highlighting the tight interaction between action of the user and the coach. We will demonstrate different parameters of our coach's coaching strategy during these virtual training sessions.

Acknowledgments

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence EXC 277 in Cognitive Interaction Technology (CITEC).

References

Iwan de Kok, Julian Hough, Cornelia Frank, David Schlangen, and Stefan Kopp. 2014. Dialogue structure of coaching sessions. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue*, pages 167–169, Edinburgh.

Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

Julian Hough, Iwan de Kok, David Schlangen, and Stefan Kopp. 2015. Timing and grounding in motor skill coaching interaction: Consequences for the information state. In *Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue*.

Herwin van Welbergen, Ramin Yaghoubzadeh, and Stefan Kopp. 2014. AsapRealizer 2.0 : The Next Steps in Fluent Behavior Realization for ECAs. In *Intelligent Virtual Agents*, pages 449–462.

Non-Sentential Utterances in Dialogue: Experiments in classification and interpretation

Paolo Dragone
Sapienza University of Rome (Italy)
`dragone.paolo@gmail.com`

Pierre Lison
University of Oslo (Norway)
`plison@ifi.uio.no`

Abstract

We present two ongoing experiments related to the classification and interpretation of non-sentential utterances (NSUs). Extending the work of Fernández et al. (2007), we first show that the classification performance of NSUs can be improved through the combination of new linguistic features and active learning techniques. We also describe a new, hybrid approach to the semantic interpretation of NSUs based on probabilistic rules.

1 Introduction

In dialogue, utterances do not always take the form of complete, well-formed sentences. Many utterances – often called *non-sentential utterances*, or NSUs for short – are indeed fragmentary and lack an overt predicate, as in the following examples from the British National Corpus:

A: How do you actually feel about that?
B: **Not too happy.** [BNC: JK8 168-169]

A: They wouldn't do it, no.
B: **Why?** [BNC: H5H 202-203]

A: [...] then across from there to there.
B: **From side to side.** [BNC: HDH 377-378]

Although these types of NSUs are extremely common, their semantic content is often difficult to extract automatically. NSUs are indeed intrinsically dependent on the dialogue context – for instance, the meaning of ”why” in the example above is impossible to decipher without knowing the statement that precedes it.

We report here on two ongoing experiments. The first experiment focuses on the automatic classification of NSUs according to the taxonomy of Fernández et al. (2007), while the second experiment develops a new approach to the semantic interpretation of NSUs using the probabilistic rules formalism developed by Lison (2015)

2 Classifying NSUs

Non-sentential utterances can serve several types of pragmatic functions, such as providing feedback, asking for clarifications, answering questions or correcting/extending previous utterances.

Fernández et al. (2007) provide a taxonomy of NSUs based on 15 classes as well as a small corpus of annotated NSUs extracted from dialogue transcripts of the British National Corpus. They also present classification experiments using the above-mentioned corpus and taxonomy. We extend their approach through a combination of feature engineering and semi-supervised learning. Semi-supervised learning is used to cope with the scarcity of labelled data for this task. This lack of sufficient training data is especially problematic due to the strong class imbalance between the NSU classes. Furthermore, the most infrequent classes are often the most difficult ones to discriminate. Fortunately, the BNC also contains a large amount of unlabelled NSUs that can be extracted from the raw dialogue transcripts using simple heuristics (syntactic patterns to select utterances that are most likely non-sentential).

One particular technique that we employed in this empirical study is Active Learning. The objective of Active Learning (AL) is to interactively query the user to annotate novel data by selecting the most informative instances (that is, the ones that are most difficult to classify) and avoiding redundant ones.[1] In practice, we applied the active learning algorithm to extract and annotate 100 new instances of NSUs, which were subsequently added to the existing training data.

---

[1] We used the Java library JCLAL for this purpose, cf. http://sourceforge.net/projects/jclal/.

In order to determine the baseline for our study, we replicated the classification experiment described in Fernández et al. (2007) using the same feature set. This initial set comprised a total of 9 linguistic features extracted from the NSU and its antecedent. We then developed an extended feature set, adding 23 new syntactic and similarity features on top of the ones used in the baseline. Weka's SMO package (based on SVMs) was used to train the classifiers for all experiments.[2]

The empirical results were extracted through 10-fold cross-validation (for the active learning case, the newly annotated instances were added to the training set of each fold). The results demonstrate that the above approach is able to provide modest but significant improvements over the baseline, as illustrated in Table 1. Using a paired t-test with a 95% confidence interval between the baseline and the final result, the improvement in classification accuracy is statistically significant with a p-value of 6.9×10^{-3}.

Experimental setting	**Accuracy**
Train-set (initial features)	0.881
Train-set (extended features)	0.899
Train-set + AL (initial features)	0.883
Train-set + AL (extended features)	**0.907**

Table 1: Summary of the classification accuracy for the baseline and new approach.

The evaluation results illustrate that the active learning approach is only beneficial when combined with the extended (more informative) feature set, while it does not provide any significant improvement on the set of baseline features.

Our experiments demonstrate the potential of the combination of linguistically-informed features and larger amounts of training data for the classification of non-sentential utterances. Of special interest would be the annotation and analysis of NSUs in other dialogue domains than the ones covered in the current corpus.

3 Interpreting NSUs

Non-sentential utterances cannot be interpreted in isolation from their surrounding context. As argued by e.g. (Fernández, 2006; Ginzburg, 2012), NSUs are best described in terms of *update rules* on the current dialogue state. Their framework is however purely logic-based, making it difficult to account for the fact that many state variables are only partially observed (due to e.g. imperfect understanding of the dialogue and its context).

To remedy this shortcoming, we are currently rewriting the update rules for NSUs detailed in (Ginzburg, 2012) using the probabilistic rules formalism of (Lison, 2015). Probabilistic rules indeed share many commonalities with Ginzburg's framework, as both approaches rely on update rules expressed in terms of conditions and effects operating on a rich dialogue state. However, probabilistic rules can also operate on uncertain (probabilistic) knowledge, making them more robust than traditional logical rules.

We are using the OpenDial toolkit[3] to implement the above approach. Crucially, the approach integrates in its pipeline the classifier presented in the previous section in order to derive the most likely class for each NSU. We plan to use a portion of the COMMUNICATOR corpus (Walker et al., 2001) to evaluate the performance of the interpretation rules on real-world dialogues.

4 Conclusion

This abstract presented two ongoing experiments related to the automatic processing of non-sentential utterances in dialogue. The first experiment shows how the use of more expressive linguistic features and active learning can improve the classification accuracy of NSUs. The second experiment focuses on the robust interpretation of NSUs in context based on probabilistic rules.

References

R. Fernández, J. Ginzburg, and S. Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

R. Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College London.

J. Ginzburg. 2012. *The Interactive Stance*. Oxford University Press.

P. Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232 – 255.

M. Walker, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, S. Seneff, and D. Stallard. 2001. DARPA Communicator dialog travel planning systems: The June 2000 data collection. In *Proceedings of EUROSPEECH*, pages 1371–1374.

---

[2] http://www.cs.waikato.ac.nz/ml/weka/

[3] http://opendial-toolkit.net

DS-TTR: An incremental, semantic, contextual parser for dialogue

Arash Eshghi
Interaction Lab
Heriot-Watt University
`a.eshghi@hw.ac.uk`

Abstract

We describe a demo to be given at the conference, of the DS-TTR dialogue parser[1]. We will show how the DS-TTR semantic context is updated in real time as dialogues are parsed incrementally, covering a variety of contextual phenomena.

1 Introduction

Language processing in dialogue is incremental and highly contextual. Dialogue is replete with fragments, ellipsis, incomplete sentences, add-ons, barge-ins, false starts, and repair (see dialogues in Fig. 1). This has had the consequence that traditional models of syntax and semantics, based strictly around the notion of a sentence have had very little success in handling dialogue phenomena, and often just put them to one side as instances of defective performance or disfluency. Although in the last decade or so, various researchers have attempted to come up with general, scalable models of semantic/contextual processing in dialogue (pioneered by the work of the likes of Ginzburg, Cooper, Traum and others (Ginzburg, 2012; Traum and Larsson, 2003)), they are hardly ever used in working, end-to-end, dialogue systems. In these existing systems, the Natural Language Understanding and Generation components are almost invariably shallow, based on pattern-matching, statistical methods, or templates, and they are highly domain-specific, thus rendering them of little or no use in a new dialogue domain. Apart from the highly domain-specific nature of meaning in general, this status quo seems to be due the apparent messiness of dialogue, as noted above, leading dialogue systems developers to use shallow statistical methods to achieve some degree of robustness in their end-to-end systems: the existing dialogue processing models alluded to above are too restrictive.

What is needed is a semantic parser/generator that is wide-coverage, capable of processing natural dialogue with all its seeming messiness; and producing domain-general, deep, re-usable semantic and contextual representations of dialogue. In what follows, we describe a working dialogue parser which is close to satisfying these properties. This is an implementation of Dynamic Syntax and Type Theory with Records (DS-TTR, (Kempson et al., 2001; Eshghi et al., 2012)), which has been in development over the past 10 or so years, showing its applicability to modelling a wide range of dialogue phenomena, including self-repair (SR) (Hough, 2015), ellipsis (Kempson et al., 2015), short-answers (SA), clarification interaction (CR), corrections (COR), split-utterances (SU) and backchannels (ACK) (see Fig. 1 for examples). It is this parser that we aim to demo at the SemDial conference, showing examples of how it handles various dialogue phenomena in real time.

2 The DS-TTR parser/generator

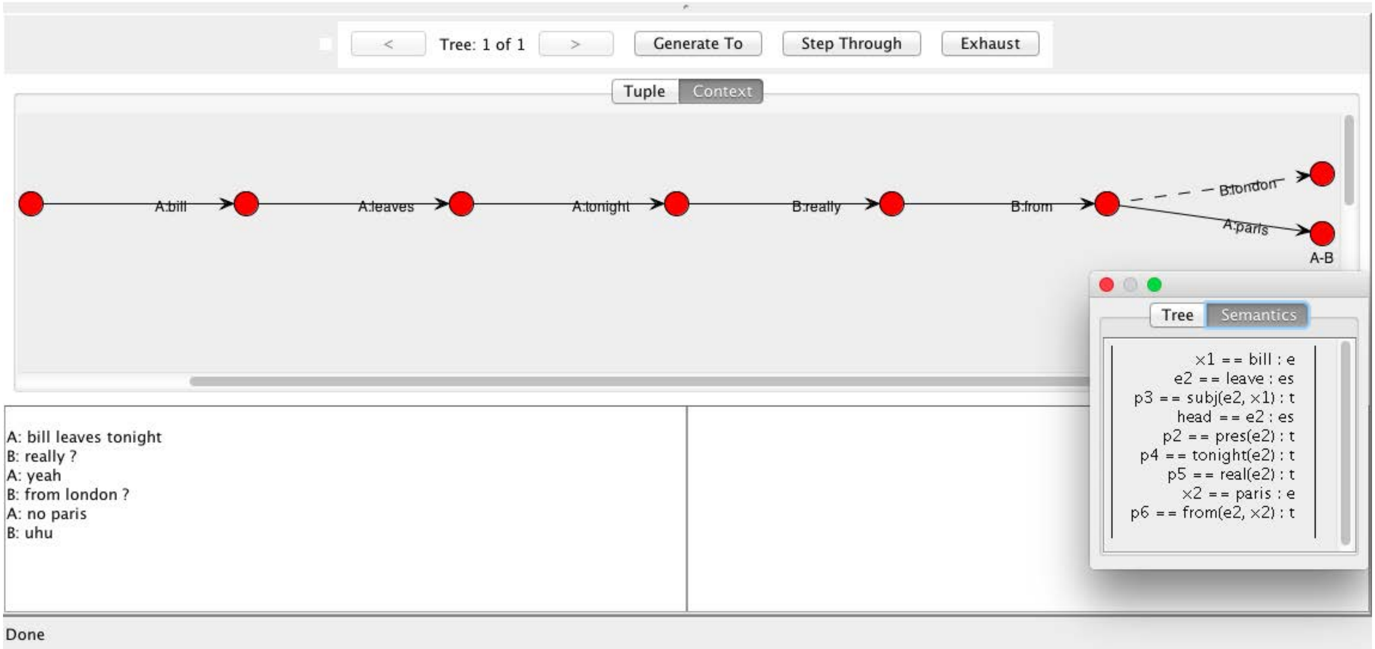
DS-TTR is an action-based parser/generator, based around the Dynamic Syntax (DS) grammar framework (Kempson et al., 2001) and Type Theory with Records (TTR, Cooper (2005)) which models the word-by-word incremental, *semantic* processing of linguistic input without recognising an independent level of syntactic representation. In DS, dialogue is modelled as the interactive and incremental construction of contextual and semantic representations. In DS-TTR, words are seen as contextual updates with context being based on the parsed search graph, a Directed-Acyclic Graph (DAG), encoding not only the fine-grained semantic contents that is jointly constructed, but also the steps (actions/words) that go on to build them

---

[1] Downloadable from: `http://sourceforge.net/projects/dylan/`. Soon to be ported to: `https://bitbucket.org/dylandialoguesystem`

(1)	(2)	(3)
A: Who did you meet yesterday?	A: Yesterday, I finally cooked uhh	A: Bill arrives tonight
B: Arash [SA]	B: What? [CR,SU]	B: Really?
A: The guy from your group? [CR]	A: Stew, uhh, Beef Stew [SR,SU]	A: Yeah
B: no, my cousin [COR]	B: with carrots? [SU]	B: From London? [SU]
A: right [ACC]. I think I have met him.	A: yeah [ACC]	A: no, Paris [Cor]
		B: uhh okay. [Ack]

Figure 1: To be demoed: Example dialogues parsable by DS-TTR



```
A: bill leaves tonight
B: really ?
A: yeah
B: from london ?
A: no paris
B: uhu
```

```
   x1 == bill : e
  e2 == leave : es
p3 == subj(e2, x1) : t
    head == e2 : es
  p2 == pres(e2) : t
p4 == tonight(e2) : t
  p5 == real(e2) : t
     x2 == paris : e
p6 == from(e2, x2) : t
```

Figure 2: Parser Screen Shot: Parsing dialogue (3), Fig. 1

(see Fig 2). Eshghi et al. (2015) show how this word-by-word contextual update process can be achieved using only the existing core mechanisms of the grammar, and capturing updates arising from feedback/grounding phenomena (backchannels and CRs) without recourse to higher level pragmatic inference, or dialogue acts.

Nodes on the context DAG are semantic representations; and edges, words indexed to speaker, i.e. semantic updates (see Fig. 2) - note that only the currently active edges and nodes are shown here: the underlying parse search DAG is much bigger than this with many more branches corresponding to parsing ambiguity. Pointers on the DAG mark nodes where each participant has given evidence of acceptance for reaching (see Eshghi et al. (2015) for details): the A-B below the final node in Fig. 2 means pointers for speakers A and B are both convergent on that node, and thus that the semantic content at that node - the TTR record type below it in the small separate window - is grounded. The branching at the end is the result of the rejection+correction (“no Paris”).

References

Cooper, R. (2005). Records and record types in semantic theory. *Journal of Logic and Computation 15*(2), 99–112.

Eshghi, A., J. Hough, M. Purver, R. Kempson, and E. Gregoromichelaki (2012). Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin (Eds.), *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, Volume 19 of *Tributes*, pp. 325–349. London: College Publications.

Eshghi, A., C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver (2015). Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguisitics.

Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Hough, J. (2015). *Modelling Incremental Self-Repair Processing in Dialogue*. Ph. D. thesis, Queen Mary University of London.

Kempson, R., R. Cann, A. Eshghi, E. Gregoromichelaki, and M. Purver (2015). Ellipsis. In S. Lappin and C. Fox (Eds.), *The Handbook of Contemporary Semantics*. Wiley-Blackwell.

Kempson, R., W. Meyer-Viol, and D. Gabbay (2001). *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.

Traum, D. and S. Larsson (2003). The information state approach to dialogue management. In Smith and Kuppevelt (Eds.), *Current and New Directions in Discourse & Dialogue*, pp. 325–353. Kluwer Academic Publishers.

What could mean *interaction* in natural language and how could it be useful ?

Christophe Fouqueré
CNRS-LIPN / Université Paris 13
`christophe.fouquere@univ-paris13.fr`

Myriam Quatrini
CNRS-I2M / Aix-Marseille Université
`myriam.quatrini@univ-amu.fr`

What could mean interaction in natural language. The notion of Interaction, which is central in different fields, from Computer Sciences to Conversational Analysis, seems to be a same term amounting to rather distinct processes. Nevertheless, we think that the same, even if abstract, concept of interaction should underlie its incarnations in different disciplines. In *Ludics*, a logical theory developed by J.Y. Girard (2001), such an abstract concept of interaction is available. We postulate that this formal approach may help us to better understand what is interaction in natural language, and therefore that some language phenomenons may be better grasped and manipulated by means of a modeling based on such conceptual considerations.

In Ludics, there is a unique primitive concept: interaction, acting according to two modes. The *closed* mode is the process of communication itself, the *open* mode accounts for the transformation that this communication process induces on contexts. We may consider that in natural language also, interaction is a common concept subsuming two modes. With the communication mode, elements of language are produced and received by interlocutors during a dialogue. With the composition mode, elements of language are composed together to produce either more elaborated elements of language or to update knowledges and commitments. Therefore, based on the Ludics theoretical frame, we proposed in (Lecomte and Quatrini, 2011; Fouqueré and Quatrini, 2013; Fouqueré and Quatrini, 2012) a dialogue modeling that accounts for both aspects of interaction: communication and computation. Our model of dialogue is organized in two levels. At the first level, called *surface of dialogues*, a dialogue is represented by an interaction between two trees, each of them is the dialogue seen from the viewpoint of one interlocutor. More precisely, each turn of speech is a sequence of dialogue acts, where each dialogue act is represented twice: once positively inside the tree associated to the speaker who produces it, and once negatively inside the tree associated to her addressee. Therefore, each turn of speech gives rise to a part of both trees growing bottom/up. At a second level, knowledges and commitments as well as linguistic elements used to build utterances are stored in two cognitive bases, each one respectively associated to each interlocutor.

Dialogical contributions such as questions, answers and concessions. Even if the types of such speech acts are at first departed by the goals and the intentions of an interlocutor during a dialogue, the inspection that our modeling enables retains more primitive features. Question and negation are not really distinguishable according to their effects on the structure of dialogues, both are particular cases of a general speech act we may call "request for justification". Its main feature, formalized at the surface of dialogue, is to be a unique dialogue act creating a unique new address where the addressee is invited to anchor her development. On another side, question and concession are very close according to their effect on cognitive bases. When she asks a question, a speaker not only formulates it, but she is ready to receive and register an answer. In the cognitive base of the speaker, the tree associated with the question contains not only the dialogue act corresponding to the formulation of the question but also the tools to *compute* the reception and the registration of possible answers by means of a copycat strategy. The argument of the function is the answer to the question, also represented by a tree. The application of such a function to its argument gives rise to an execution: an interaction between the two trees. After the interaction, the cognitive base is augmented with the information given by the answer. To sum up, in the cognitive base, the question is associated to a tree which enables an updating. Moreover, we may remark that the effect of concession in cognitive bases is similar: when an interlocutor concedes a position claimed by her addressee, she records this position in her own cognitive base, still using a tree which enables to copy such a position and record it.

Grasping cognitive processes as computation. D. Prawitz (2007) studies the elements that determine the validity of inferences. In particular, he shows that the Modus Ponens rule is insufficient for taking into account the cognitive process at stake when an addressee is convinced by an argumentation. Instead it is the phenomenon of cut elimination which accounts correctly for what is responsible of the conviction. For D. Prawitz, the cognitive process requires a proof of one premise followed by the deductive extraction from this premise towards a conclusion. By this way, the addressee of an argumentation is obliged to accept an inference, if she stays rational. Our dialogue modeling follows and even more goes further Prawitz analysis. According to the theoretical framework on which our modelization is based, a proposition is denoted by the set of its justifications, whereas classically a proposition is formalized as a simple logical formula. In the same way, a “claim”, a “thesis”, a “belief” on which a protagonist commits herself during a controversy, is denoted by a sequence of arguments in a proof-like style. Such sequences of arguments make explicit the process according to which the protagonist is convinced by the validity of her commitments. It is worth noticing the two following points:

- Such a justification is formally a cut-free proof. It is the trace of the thought process which achieves the conviction about a proposition (close mode).
- Cut-free proofs at stake may interact: This process (open mode) yields a new cut-free proof that represents the new knowledge.

References

Christophe Fouqueré and Myriam Quatrini. 2012. Ludics and Natural Language: First Approaches. In *Logical Aspects of Computational Linguistics*. Folli-LNAI, Springer.

Christophe Fouqueré and Myriam Quatrini. 2013. Argumentation and inference: A unified approach. *Baltic International Yearbook of Cognition, Logic and Communication*, 8(4):1–41.

Jean-Yves Girard. 2001. Locus solum: From the rules of logic to the logic of rules. *Mathematical Structures in Computer Science*, 11(3):301–506.

Alain Lecomte and Myriam Quatrini. 2011. Figures of Dialogue: a View from Ludics. *Synthese*, 183:59–85.

Dag Prawitz. 2007. Validity of inferences. In *”The 2nd Launer Symposium on Analytical Philosophy on the Occasion of the Presentation of the Launer Prize, Dagfinn Fllesdal”, in Bern.*

Towards the Automatic Extraction of Corrective Feedback in Child-Adult Dialogue

Sarah Hiller and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
`sarah.hiller@posteo.de` `raquel.fernandez@uva.nl`

Abstract

We present the first steps of a project that aims to investigate the effects of *corrective feedback* on language acquisition. We propose a methodology for the automatic extraction of instances of corrective feedback from child-adult dialogue corpora and discuss our plans for a data-driven investigation of this phenomenon.

1 Corrective Feedback as Negative Input

Children learn language in interaction with proficient speakers around them. This naturally allows them to be exposed to *positive input*, i.e., grammatical utterances in context. It is however a matter of debate whether children receive any form of *negative input*, i.e., corrections or indications that point out the mistakes in their utterances. Researchers such as Brown and Hanlon (1970) have pointed out that caregivers' explicit approval or disapproval of a child utterance is not contingent on its grammaticality, but rather on the appropriateness of its meaning. This has for a long time been taken to show that children do not receive any negative input (see e.g. Chouinard and Clark (2003) for a discussion of these issues).

This judgement, however, can arguably be considered premature since explicit disapprovals are not the only possible means of highlighting grammatical errors. Indeed, Brown and Hanlon (1970) already noted that "[r]epeats of ill-formed utterances usually contained corrections and so could be instructive". In a similar trend, Chouinard and Clark (2003) and Saxton (2010) make a case arguing that corrected repetitions of children's ungrammatical utterances constitute a form of negative input. In addition to pointing towards an error, this strategy also presents the correct form, as shown in the following example from CHILDES (MacWhinney, 2000) by 2-year-old Lara:

(1) CHI: what about kiss ?
DAD: what about a kiss ?

We refer to this kind of child-adult exchanges as *corrective feedback*. As can be seen in (1), they are characterised by a child utterance with some grammatical anomaly followed by an adult response that repeats part of the child's utterance and modifies it, thereby offering a grammatically correct counterpart to the child's error.

Different accounts have been put forward to explain what triggers this kind of adult responses. For instance, Chouinard and Clark (2003) consider that they arise as a side effect of parents checking up on the *meaning* of children's utterances, while Saxton et al. (2005) claim that it is the *form* of the child utterance that is directly at issue. In any case, all approaches agree that such responses create a contrast that may act as a correction.

2 Manual Annotation

In order to study the properties of corrective feedback and the effect it may have on language learning, we are interested in developing data-driven methods that allow us to automatically extract instances of corrective feedback from dialogue transcripts at a large scale. For our long-term investigation, we consider all the CHILDES transcripts from children with no impairments for which there is data for a minimum period of 1 year with at least five dialogues per year. We select only those dialogues with a minimum length of 100 utterances (and at least 50 child utterances) where the child speech has a minimum mean length of utterance (MLU) of 2 words.

From this set, we select 16 files from 4 randomly selected children (4 files per child) for detailed analysis. To arrive at an automatic extraction algorithm, we first apply a very simple heuristics to obtain child-adult utterance pairs that are candidates for corrective feedback: we select all

Linguistic level		Type
Syntax (49.3%):	Subject, Verb, Object, other	Omission (81.3%)
Noun Morphology (3.3%):	Possessive -'s, Regular Plural -'s, Irregular Plural, other	Addition (3.0%)
Verb Morphology (3.7%):	3rd Person singular -s, Regular past -ed, Irregular past, other	Substitution (14.6%)
Unbound Morphology (31.5%):	Determiner, Preposition, Auxiliary verb, Present progressive, other	Other (1.0%)
Other (12.2%):	–	

Table 1: Distribution of errors according to linguistic levels and types, together with their frequency counts in exchanges containing corrective feedback (307 instances in total).

exchanges where there is partial overlap between the child and the adult utterances and where the child's utterance includes at least two word types (a total of 2072). We then manually annotate these instances to filter out false positives. Those exchanges identified as corrective feedback are additionally annotated with information regarding the error being corrected: the linguistic level at which the error occurs (based on Saxton et al. (2005)) and the type of error (based on Sokolov (1993)) – see Table 1. For example, the exchange in (1) would be annotated as *Unbound Morphology: Determiner – Omission*. To compute inter-annotator agreement for the *corrective feedback* identification task, 350 instances from two different files were annotated by two coders, obtaining a Cohen's κ of 0.77.

3 Results and Next Steps

Next we describe the results obtained so far. Of the 2072 pairs of utterances annotated, 14.8% are identified as instances of corrective feedback. Note that this number is not representative of how many *errors* are met with corrections, since the candidate utterance pairs also contain correct child utterances. The frequency distribution of error types amongst the exchanges tagged as corrective feedback is shown in Table 1. As can be seen, most errors are of syntactic nature (49.3%) and concern omissions (81.3% over all linguistic levels). The high number of omission errors is perhaps not surprising in child language, given the comparably low MLU. Corrective feedback decreases over time, as children make less errors – see Figure 1, which also shows that frequency of corrective feedback varies largely between children. This will be useful for the comparative investigation of its effects on language learning.

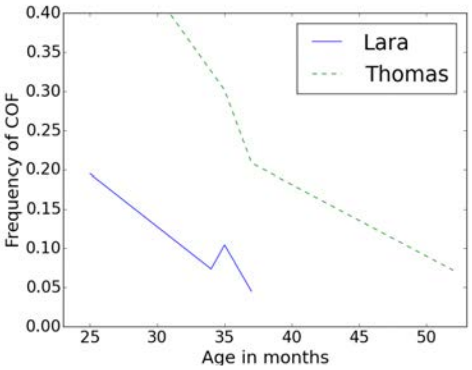


Figure 1: Frequency of corrective feedback against child age for two children. Pearson coefficient for Lara is -0.96, for Thomas -0.95.

Our next step is the development of algorithms for the automatic extraction of corrective feedback. We will first extract a set of features representative of the instances annotated as corrective feedback, including level of overlap, syntactic dependency information, and semantic distance, and train a supervised machine learning classifier. Once we have a sufficiently reliable extraction method, we will investigate to what extent corrective feedback is helpful in language acquisition. For this we will compare adult constructions which often occur as corrective feedback to those which occur in non-contingent environments.

References

Roger Brown and Camille Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In John R. Hayes, editor, *Cognition and the Development of Language*. John Wiley & Sons, Inc.

Michelle M. Chouinard and Eve V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(03):637 – 669.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, 3 edition.

Matthew Saxton, Phillip Backley, and Clare Gallaway. 2005. Negative input for grammatical errors: effects after a lag of 12 weeks. *Journal of Child Language*, 32(03):643 – 672.

Matthew Saxton. 2010. *Child language. Acquisition and Development*. SAGE Publications.

Jeffrey L. Sokolov. 1993. A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, 29(6):1008 – 1023.

Spot the Difference
- A dialog system to explore turn-taking in an interactive setting

Anna Hjalmarsson
Department of Speech, Music and
Hearing, KTH, Stockholm, Sweden

annahj@kth.se

Margaret Zellers
Department of Speech, Music and
Hearing, KTH, Stockholm, Sweden

zellers@kth.se

Abstract

Much research has been devoted to understanding the principles that control the flow of dialog contributions between speakers in dialog. This demonstration paper describes a dialog system that was developed as test-bed to experiment with turn-taking aspects in an interactive setting.

1 Introduction

Over the years, much effort has been devoted to understanding the principles that control the flow of dialog contributions between speakers (see for example Sacks et al., 1974). The motivation for this research includes both the desire to understand the underlying mechanisms of human communication as well as to build dialog systems with more sophisticated turn-taking capabilities. A first step towards increased understanding of these phenomena has been to identify behaviors that correlate with speaker changes in human-human dialog (see for example Duncan, 1972). One approach to further understand how these behaviors influence listeners' expectations of a speaker change is to study listeners' expectations of who will speak next in an off-line setting where subjects listen to pre-recorded dialog excerpts (Hjalmarsson, 2011 and Zellers, 2013). However, in order to understand to what extent the target behaviors actually influence listeners' turn-taking decisions, these behaviors needs to be explored in an interactive setting. The aim of the dialog system presented in this demonstration paper is to serve as a test-bed for such experimentation. An advantage of using a dialog system to do this is that a system's behavior, as opposed to a human's behavior, can easily be controlled. Furthermore, a dialog system is also suitable for studies that aim to identify human behaviors that can be used to regulate turn-taking in human-machine interaction.

The paper is structured as follows. In section 2, we will present the motivation and theoretical background of an initial planned study and in section 3, we will present the domain and implementation of the dialog system that we will use in this research.

2 Timing in utterance generation

Most of today's dialog systems have no strategies to adjust the timing of speech to the local dialog context. Utterances are produced as whole units as soon as they become available to the speech generator, and the timing of individual speech segments is typically based on a shallow syntactic analysis of the isolated utterance. However, dialog systems that use incremental models for processing (Schlangen & Skantze, 2011) process utterances in smaller sub-segments in a way that is more similar to human speech processing. Such incremental speech processing opens up for more fine-grained generation of utterances where small variations in the system's output can be used to accommodate the semantic and pragmatic dialog context. Analyses of human-human dialog data suggest that the temporal flow of speech has several important structural functions (cf. Goldman-Eisler, 1972). The timing of different speech events – a phoneme, a prolonged syllable or a pause – in conversation affects listeners' perception of an utterance and is influenced by the dialog context (Zellner, 1994).

In a recent series of articles (Skantze & Hjalmarsson, 2013 and Skantze et al., 2014), we have explored how the preceding context affects users' reactions to temporary silences in the system's speech. The aim of the system presented here is to serve as a testbed for pursuing this research in the setting of fully functional dialog system. In an initial experimental study, we will explore how various non-lexical behaviors, such as variation in pitch and duration as well as inhalations and fillers (e.g. “eh” and “ehm”) affect users’ turn-taking decisions when these behaviors are followed by silence.

3 The Spot the Difference system

The domain that was chosen for the dialog system is similar to the frequently used map-task domain (Anderson et al., 1991). However, instead of identifying differences between maps, the players’ task is to identify differences between two versions of a picture (see Figure 1).

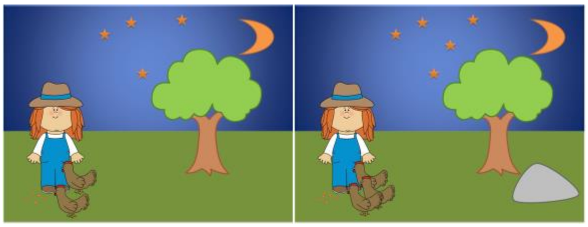
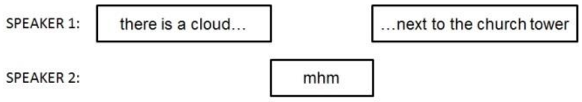


Figure 1: Two versions of a scene in the system.

In this domain, nominal phrases of various complexity are used to refer to objects, and whether it is appropriate to take the turn or not is often ambiguous when relying on lexical context alone (see the dialog example in Figure 2). This makes the domain suitable for experimenting with non-lexical turn-taking cues.



SPEAKER 1: there is a cloud... ...next to the church tower

SPEAKER 2: mhm

Figure 2: Human-human dialog excerpt.

3.1 System implementation and setup

The dialog system was implemented using IrisTK (Skantze & Al Moubayed, 2012), a framework for building multimodal conversational systems, and the GUI was implemented in Java. For automatic speech recognition and end-of-speech-detection, we use an off-the-shelf speech recognizer, and for speech synthesis, we use the CereVoice system developed by CereProc[1]..In order to explore the effect of mid-utterance pauses, the system’s utterances are realized in utterance segments with short silences in-between. As the aim of the experiment is to explore effect of non-lexical behaviors, all utterance segments are semantically complete.

Acknowledgments

This work was supported by the Swedish Research Council (VR) project Classifying and deploying pauses for flow control in conversational systems (2011-6152).

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351-366.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15*, 103-113.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1), 23-35.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Schlangen, D., & Skantze, G. (2011). A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse, 2*(1), 83-111.

Skantze, G., & Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA.

Skantze, G., & Hjalmarsson, A. (2013). Towards Incremental Speech Generation in Conversational Systems. *Computer Speech & Language, 27*(1), 243-262.

Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication, 65*, 50-66.

Zellers, M. (2013) Pitch and lengthening as cues to turn transition in Swedish. Proceedings of 14th Interspeech, Lyon, France, 248-252.

Zellner, B. (1994). Pauses and the Temporal Structure of Speech. In Keller, E. (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41-62). Chichester: Wiley.

---

[1] http://www.cereproc.com

Blinking as addressee feedback in face-to-face dialogue?

Paul Hömke, Judith Holler, Stephen C. Levinson
Language and Cognition Department,
Max Planck Institute for Psycholinguistics,
Nijmegen, The Netherlands
`paul.hoemke@mpi.nl, judith.holler@mpi.nl,`
`stephen.levinson@mpi.nl`

Abstract

Does addressee blinking function as a type of visual feedback in face-to-face dialogue? Preliminary quantitative analyses reveal that in a corpus of spontaneous Dutch dialogue the majority of addressee blinks was timed like types of addressee behavior with clear feedback functions, namely to the end of speaking units within long turns. Preliminary qualitative analyses reveal that long addressee blinks (>=410ms) were produced especially after speakers' self-repairs. This suggests that addressee blinking is closely linked to the structure of speakers' turns, and that in addition to potential cognitive functions, especially long addressee blinks may function as a social feedback signal of understanding.

1 Introduction

In face-to-face dialogue, the addressee provides vocal and visual feedback while the speaker is speaking (e.g., *hm-hm,* head nod; Yngve, 1970). Is eye blinking, too, a type of visual addressee feedback? People blink more often than necessary for wetting their eyes and they tend to blink after they think. That is, blink rate increases with low cognitive load and decreases with high cognitive load (Siegle et al., 2008). But blinking has also been linked to social functions. Comparing blink rates across different activities the highest blink rate was found in conversation, and in non-human primates, blink rate is positively correlated with group size, a measure of social complexity (Tada et al., 2013). These findings suggest that in addition to peripheral physiological and central cognitive functions, human blinking may have a social-communicative function.

Within conversation, blinking is the most frequent facial action and, in American Sign Language, addressees use blinks to signal understanding (Sultan, 2004). Sultan argued that addressee blinking might have developed a feedback function in sign language because of the need to control blinking to minimize visual information loss.

In the present study, we hypothesized that addressee blinking may have a feedback function in spoken Dutch, too, because many spoken languages also rely heavily on the visual channel, at least in face-to-face contexts. If this was true, one should expect addressee blinks to be timed like other types of addressee feedback, namely to the end of speakers' syntactically, prosodically, and pragmatically complete units within turns. If our hypothesis was wrong one would expect addressee blinks to be distributed randomly and irrespective of the communicative context.

2 Method

To address this question, we built an audio-video corpus of informal, spontaneous Dutch face-to-face dialogue (10 dyads: 4 female-female, 4 female-male, 2 male-male; 18-68 years) and—focusing on multi-unit turns—we measured the temporal distance of each addressee blink onset to the closest end of a speaker's syntactically, prosodically, and pragmatically complete unit. Here is an example of a multi-unit turn (ends of speaker units are marked by a |): *If you both become happy | this is more important | than that your home remains as it was* |. Addressee blinks were detected semi-automatically using a motion tracking software (Xiong & De la Torre, 2013) combined with manual coding.

3 Results

Preliminary quantitative analyses revealed that the majority of addressee blinks occurred very close to the end of speaker units (see **Figure 1**).

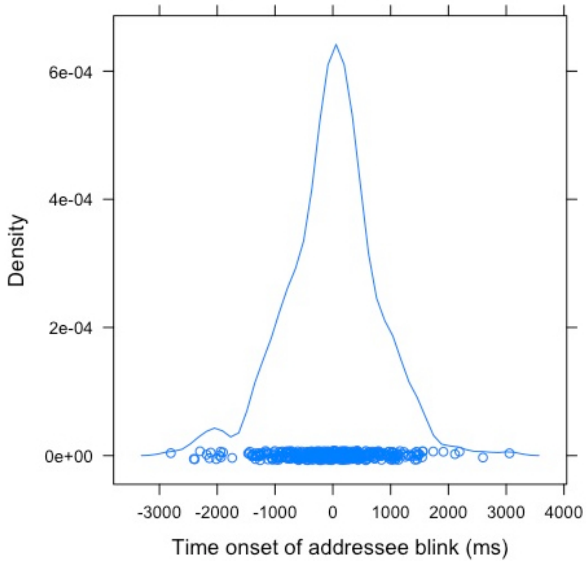


Figure 1. Addressees' blink onset (N=411) measured to the closest speaker unit end. The zero point on the x-axis marks the end of the speaker unit, the peak of the density distribution the estimate of the mode, and dots represent individual data points.

Preliminary qualitative analyses revealed that long addressee blinks (>=410ms) were especially produced after speakers' self-repairs. Here is an example (translated from Dutch): A (on the left) asked B (on the right): *Did he send you a letter in response?* B then answered and added a self-repair (underlined): *He sent me a message afterwards - well it wasn't a letter but on Whatsapp a long message*. After B's self-repair, as soon as she looked back at her addressee, the addressee responded with a long blink (and a head nod; see **Figure 2**).

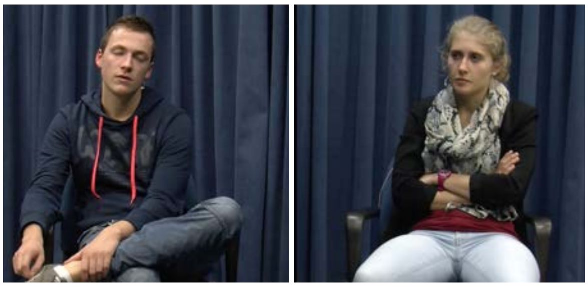


Figure 2. Example of a long addressee blink (on the left) following a speaker's self-repair. Note that this image is taken from a split-screen recording and that participants were facing each other in actuality.

4 Discussion and Conclusion

While all blinks also lubricate the cornea, addressee blinks were produced too frequently to serve solely this physiological function. The fact that the majority of addressee blinks was timed to unit ends is consistent with a cognitive interpretation of blinking (Siegle et al, 2008) because it may reflect addressees' relative decrease in cognitive load. But the results are also consistent with a social interpretation of blinking because speakers tend to visually monitor addressees for feedback at unit ends, and in addition, especially long addressee blinks following speakers' self-repairs seem to function as a social signal of successful grounding (Clark & Brennan, 1991). Cognitive and social functions of addressee blinks are not mutually exclusive, of course. Maybe the cognitive function underlies and evolutionarily preceded any potential social function. Perhaps blinking as a symptom of momentary low cognitive load has been co-opted for communicative purposes, so that it is now (also) used as a social signal.

Acknowledgements

We thank Binyam Gebrekidan Gebre, Mart Lubbers, and Emma Valtersson for programming support and three anonymous referees for valuable comments. This research was funded by the European Research Council (Advanced Grant #269484 INTERACT awarded to Stephen Levinson) and the Max Planck Gesellschaft.

References

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition*, *13*(1991), 127-149.

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*(5), 679-687.

Sultan, V. D. (2004). *Blinking as a Recipient Resource in American Sign Language Discourse* (Doctoral dissertation, University of California, Santa Barbara).

Tada, H., Omori, Y., Hirokawa, K., Ohira, H., & Tomonaga, M. (2013). Eye-blink behaviors in 71 species of primates. *PloS one*, *8*(5), e66018.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting* (pp. 567-578).

A Reusable Interaction Management Module:
Use case for Empathic Robotic Tutoring

Srinivasan Janarthanam, Helen Hastie,
Amol Deshmukh, and Ruth Aylett
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh

Mary Ellen Foster
School of Computing Science,
University of Glasgow, Glasgow

Abstract

We demonstrate the workings of a stochastic Interaction Management and showcase this working as part of a learning environment that includes a robotic tutor who interacts with students, helping them through a pedagogical task.

1 Introduction

We demonstrate the workings of a stochastic Interaction Management (IM) module, show-casing a use-case where this IM has been implemented as a part of a robotic tutor who can sense the user's affect and respond in an empathic manner. The IM is designed to be re-usable across interactive tasks, using a scripting language. We use an Engine-Script design approach, so that the IM can be used as part of the conversational agent as well as user simulations.

2 An Empathic Robotic Tutor

An empathic robotic tutor was designed as part of the Emote FP7 project[1] to aid students aged between 10-13 years in two different scenarios: a map reading task and a serious game on sustainable development (Deshmukh et al., 2013). The architecture of the system is shown in Figure 1.

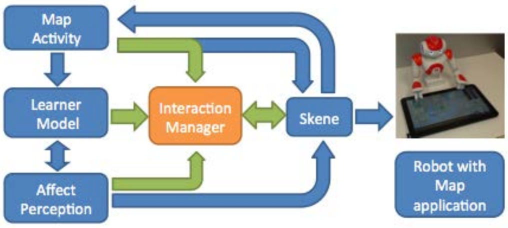


Figure 1: Architecture of the system

3 Interaction Management

The IM can be seen as the central decision making body in the architecture. It is responsible for updating and maintaining the context of the interaction and also for deciding how to respond to the input received.

The module was designed in two parts: Engine and Script. The engine is a generic implementation of the functionalities of the IM while the script presents the details of the interaction task to the engine. The major advantage of keeping the engine and the script separate is the reusability factor: the IM can be reused for other interactive tasks by simply changing the script.

3.1 IM Engine

The IM Engine implements the Information State Update approach (Larsson and Traum, 2000), extended further to include stochastic behaviours and learning capability. When presented with an input, the IM engine executes a two-step process: update context and select next action. Both steps are driven by a set of rules specified in the script. All rules to update context whose preconditions are satisfied are executed. However, for action selection rules, the IM follows one of the following approaches:

1. First Fire: Execute the rule whose precondition is satisfied first. In this approach, the order in which the rules are placed in the script file is important.

2. Collect and Select: Collect all actions whose preconditions are satisfied and select one at random.

Action selection rules can be stochastic. Within a given rule there can be several actions, each set with a probability of execution, provided the preconditions are met. In addition, the IM can be set

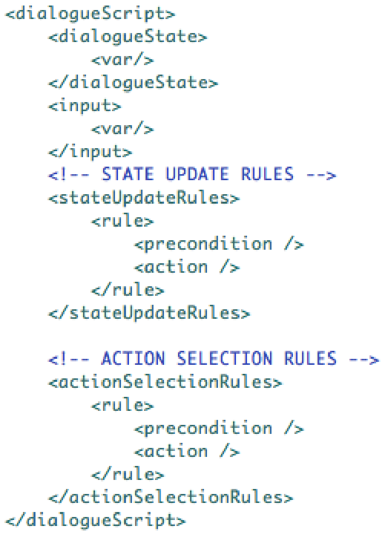
---

[1] http://emote-project.eu

to run as a Reinforcement Learning agent, to optimise its choice of actions for a given reward function, instead of randomly selecting one (i.e. in *collect and select*).

3.2 IM Script

The IM script defines context and behaviour of the Interaction Manager. It informs the IM Engine on how state update and action selection needs to be carried out for any given interactive task. The script also defines the state of the interaction, which is used to maintain the context of the conversation. The script is written in a formal language in the form of an XML document. The top level elements of the script is shown in Figure 2. These include the dialogue state and input specifications, state update and stochastic action selection rules (Figure 3).



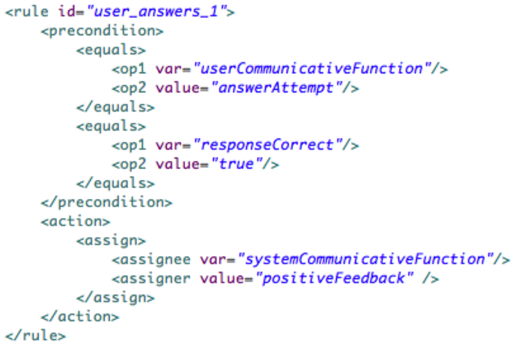
```
<dialogueScript>
    <dialogueState>
        <var/>
    </dialogueState>
    <input>
        <var/>
    </input>
    <!-- STATE UPDATE RULES -->
    <stateUpdateRules>
        <rule>
            <precondition />
            <action />
        </rule>
    </stateUpdateRules>

    <!-- ACTION SELECTION RULES -->
    <actionSelectionRules>
        <rule>
            <precondition />
            <action />
        </rule>
    </actionSelectionRules>
</dialogueScript>
```

Figure 2: Dialogue Script (Top level elements)

The IM engine can manifest both as a conversational agent as well as a simulated user by using two instances of the IM engine with different IM scripts.

4 EMOTE Tutor

We combined both empathic and pedagogical strategies in a unique and natural way in order to provide an effective learning experience using the IM script. We will demonstrate how the above IM tool was used in the context of the EMOTE empathic robotic tutor. During the demonstration, we will describe the design of the dialogue states for the two scenarios, the modelling of input and trigger events and the implementation of update and action rules. We will demonstrate how the IM was used both as part of the tutor as well as for learner simulations. We will also explain the key features of the scripting language and show how new interactive tasks can be designed and implemented using the framework.



```
<rule id="user_answers_1">
    <precondition>
        <equals>
            <op1 var="userCommunicativeFunction"/>
            <op2 value="answerAttempt"/>
        </equals>
        <equals>
            <op1 var="responseCorrect"/>
            <op2 value="true"/>
        </equals>
    </precondition>
    <action>
        <assign>
            <assignee var="systemCommunicativeFunction"/>
            <assigner value="positiveFeedback" />
        </assign>
    </action>
</rule>
```

Figure 3: Example Action Selection Rule

5 Conclusion

The IM has been evaluated in three different studies in May 2015. We propose to demonstrate the workings of the reusable stochastic Interaction Manager built to power human robot interactions. This will be done with a NAO robot and touch enabled 18” tablet running the learning scenarios. The conference delegates will be able to interact with the system and experience the empathic behaviour of the robot[2]. We also hope that the SEMDIAL community will be interested in using the IM tool as a part of their future projects.

6 Acknowledgements

This work was supported by the European Commission (EC) and was funded by the EU FP7 ICT-317923 project EMOTE. We would like to acknowledge the other Emote partners for their collaboration in developing the system described here.

References

A. Deshmukh, G. Castellano, A. Kappas, W. Barendregt, F. Nabais, A. Paiva, T. Ribeiro, I. Leite, and R. Aylett. 2013. Towards empathic artificial tutors. In *Proceedings of the 8th HRI conference*, pages 113–114.

S. Larsson and D. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323 340.

---

[2] Demo video of an early prototype of the system: `https://youtu.be/Rm7pE70KFn0`

Concern-Alignment Analysis of Consultation Dialogues

Yasuhiro Katagiri
Future University Hakodate, Japan
katagiri@fun.ac.jp

Katsuya Takanashi
Kyoto University, Japan
takanashi@kyoto-u.ac.jp

Abstract

‘Concern Alignment in Conversations’ project aims, through empirical examinations of real-life consensus-building conversations, to investigate the interrelationship between rational processes of agreement seeking and affective processes of trust management in conversational interactions. We analyzed a series of venture consultation sessions between prospective business start-up candidates and venture incubation consultants within the *concern alignment* model. We argue the concern alignment model provides us with a conceptual frame to examine consultation interaction as a process in which participants collaboratively explore the space of potential concerns to identify and examine relevant concerns to be addressed, which enable them to expand and elaborate their business proposals.

1 Concern alignment

Concern align model (Katagiri et al., 2013a; Katagiri et al., 2013b) conceptualize dialogue processes in consensus decision-making as consisting of two functional parts, concern alignment and proposal exchange, as shown in Figure 1. When a group of people engage in a conversation to find a joint course of actions among themselves on certain objectives (*issues*), they start by expressing what they deem relevant on the properties and criteria on the actions to be settled on (*concerns*). When they find that sufficient level of alignment of their concerns is attained, they proceed to propose and negotiate on concrete choice of actions (*proposals*) to form a joint action plan. Expanding on the works to establish a comprehensible set of dialogue acts (Bunt, 2006) for speech acts performed by utterances, we stipulate a set of discourse acts at the level of concern alignment in terms of functions a discourse segment perform in consensus-building, as shown in Table 1.

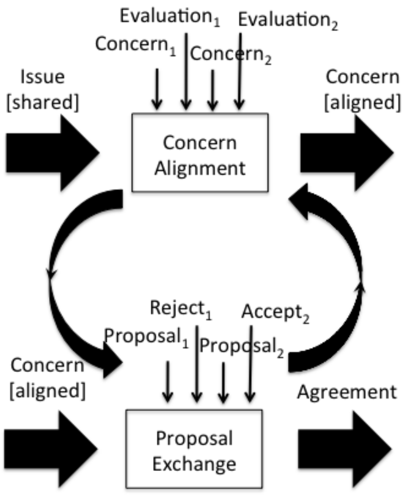


Figure 1: A concern alignment model for dialogue structures in consensus-building conversations.

2 Collaborative exploration of concern space in consultation conversations

Concern introduction as criticism to proposals: In consultation-type conversations, proposals often put on the table for discussion before relevant concerns are raised and examined. Depending on who raised those concerns following the proposal, they can work either as a support or a criticism of the proposal. Figure 2 shows an example in which the concern introduced by the consultant A, which follows the initial proposal by the business start-up candidate C, effectively works as a criticism

Table 1: Discourse acts in concern alignment

Concern alignment	
C-solicit	solicit relevant concerns from partner
C-introduce	introduce your concern
C-eval/positive	positive evaluation to introduced concern
C-eval/negative	negative evaluation to introduced concern
C-elaborate	elaborate on the concern introduced
Proposal exchange	
P-solicit	provide relevant proposal from partner
P-introduce	introduce your proposal
P-accept	provide affirmation to introduced proposal
P-reject	indicate rejection to introduced proposal
P-elaborate	modify the proposal introduced

C:	P-introduce:	*provide service to estimate market value of user skills*
A:	C-introduce	*how to justify method/criteria of estimation*
D:	P-introduce:	*provide assessment at skill category level*
A:	(ack)	
D:	P-introduce:	*leave room for variation based on peer estimation*
A:	(ack)	

Figure 2: C-introduce as criticism.

C:	P-introduce:	*web site for providing service to match up people with needs and people with skills*
B:	C-introduce	*how to find ways to attract users*
C:	(ack)	
B:	P-introduce:	*provide the matching service as* mixi *App.*
C:	(req-clarify)	*our proposal does not have the functionality?*
B:	C-introduce:	*how to find ways to attract users* /mixi *already has rich user base*
C:	(ack)	

Figure 3: P-introduce as concern foregrounding.

by presenting a potential difficulty in the proposal, which, in turn, can invite the candidate to abandon and pursue alternative proposals, or, as in this case, to elaborate on the present proposal to add details to circumvent the difficulties.

Proposal introduction as foregrounding concerns: Proposals, even when they are presented as hypothetical examples, can be used to highlight relevant concerns to be seriously entertained. Figure 3 shows an example in which an initial business proposal presented by the start-up candidate C was countered by an alternative proposal by the consultant B, which effectively focus attention to the significance of developing an idea to secure large enough user base to develop a promising business plan.

A:		*lost the grasp of what you really want to do in your business*
D:	(ack)	
A:	C-introduce:	*would you pursue ways to realize a market place for people to do whatever they want to do*
D:	(ack)	
A:	C-introduce	*Or would you pursue ways to realize a community for people to get satisfaction through their face-to-face social interactions*
D:	(ack)	

Figure 4: Parallel C-introduce for concern space exploration.

Parallel concern introduction for concern space exploration: A set of parallel concerns can be contrastingly introduced. Figure 4 shows an example in which the consultant A, after indicating his frustration on not getting a clear idea on the goals of the start-up candidates, indicated, in the form of two parallel concerns: matching for business or place for social interaction, two alternative directions they might pursue. Parallel concerns give structures to the potential space of concerns to be considered, come up with alternative lines of proposals to be pursued and force participants to make a choice among those alternatives.

Sequential patterns identified here capture strategies employed in conversations in which participants jointly push forward to explore, to motivate and to organize their thinking to eventually develop a concrete proposal to be presented and evaluated by venture capitalists.

3 Conclusions

We conducted an analysis of consultation conversations based on the *concern alignment* model. We identified several sequential organization patterns of the exchange of concerns and proposals, which successfully capture some of the strategies adopted in the process of collaborative development of consensus proposals. The notion of concern alignment provides us with a promising descriptive framework to elucidate both the processes and strategies in a wide range of consensus-building conversations.

Acknowledgments

The research reported in this paper is partially supported by Japan Society for the Promotion of Science Grants-in-aid for Scientific Research (B) 15H02752.

References

Harry Bunt. 2006. Dimensions in dialogue act annotation. In *the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Yasuharu Den, and Mika Enomoto. 2013a. Concern alignment and trust in consensus-building dialogues. *Procedia - Social and Behavioral Sciences*, 97:422–428.

Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Mika Enomoto, Yasuharu Den, and Shogo Okada. 2013b. Analysis and modeling of concern alignment in consensus-building. In *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue (SemDial2013)*, pages 202–203.

The interactive building of names

Stergios Chatzikyriakidis
LIRMM, University of Montpellier 2
`stergios.chatzikyriakidis@lirmm.fr`

Ruth Kempson
King's College London
`ruth.kempson@kcl.ac.uk`

Ronnie Cann
University of Edinburgh
`r.cann@ed.ac.uk`

1 Introduction

With the parsing and production of natural languages increasingly established as fully incremental processes, the observed interactivity of participants in developing the content of conversational dialogues becomes much less of a puzzling phenomenon (Purver et al., 2006; Gregoromichelaki et al., 2011; Howes et al., 2011; Ginzburg, 2012) In conversational dialogue, speakers and hearers can switch roles at any point so that linguistic dependencies can be split between participants at any level of the discourse including sub-sentential ones. One person provides the linguistic environment for establishing some upcoming dependency –a phrasal head, an antecedent, the source for some ellipsis– for which the other interlocutor provides the follow-up dependent element –a complement, a pronominal, the ellipsis site, etc. Here, we consider the co-construction and construal of indefinite existential terms, in which we see the same potential for distribution of the contributing expressions across more than one participant:

(1) A: She needs a a
B: mattock. For breaking up clods of earth.

(2) A: We visited
B: a friend of Granny's
A who is recovering from a post-op infection.

The goal is to argue that indefinites can be analysed, like all other natural language expressions, in terms of mechanisms which are grounded in the potential they allow for coordinative interaction, despite their quantificational nature and hence scopal properties.

2 Dynamic Syntax

We adopt the Dynamic Syntax framework as background, in which the process of constructing meanings from strings of words incrementally is central to explanations of syntactic and semantic phenomena of natural language. Underspecification of meaning-structure representations and update of these are core notions of the framework (Kempson et al., 2001; Cann et al., 2005). Both emergent content and the attendant shifting context are defined over the transition between partial representations (shown as binary branching trees), as driven by partly top down, partly bottom up processes, evolving on a word by word basis. Production and parse activities operate in tandem with reference to some current structural state in anticipation of some upcoming update. In either activity, essentially similar partial semantic trees are developed, and switch of roles is predicted to be seamless. The only difference between the activities is that whereas the parser has only a relatively weak goal to fulfill, the construction of some meaning from the linguistic input, the speaker has a more particular goal, that of the content of what she wishes to say, relative to which all construction steps have to be checked for commensurability.

Macros of action sequences triggered by words, constituting their contribution to interpretation, are a major source of the tree-growth progression. The emergent trees reflect the structure of some predicate-argument representation of content to be paired with some emergent NL string, the building of which is driven by a combination of general strategies and such lexically induced sequences of macros of actions. Quantifying expressions are taken to induce terms of the epsilon calculus, invariably of type e, denoting witness sets, as are temporal specifications, which are mapped onto sortally restricted eventuality terms, both being built up as part of the process of meaning construction. These "syntactic" mechanisms, being meta- to the representations themselves, are actions defining HOW parts of representations of content can be introduced and updated, all such growth being relative to context, itself an evolving sequence of (partial) tree structures. Reflecting compositionality of content as defined on such output trees, the individual nodes of that tree carry decorations of the (sub)-formulae of the predicate-argument formula finally derived. The update process taken to yield such a tree operates subject to a strict word-by-word incrementality. At all non-final stages of tree construction, there are open requirements that need to be satisfied. These take the form of $?X$ for any annotation X and the system defines actions that give rise to (possibly modal) expectations inducing further actions at subsequent stages of tree development. The progressive satisfaction of requirements as these get incrementally introduced yields incremental updates of some emergent structure towards some overall goal, the output tree with no requirements outstanding.

Recent work (Kempson, Fcmg) has argued that anaphora and predicate ellipsis are canonical instances of interaction in virtue of the antecedent-expression chains built up both within and across utterance boundaries, anaphorically and cataphorically. Furthermore, the mechanisms underpinning local and long-distance discontinuities are equally interactive, in displaying the same dynamic patterning allowing some underspecified parameter to be resolved: from established context; from local context emergent from the construction process; or even, given the domain-general vocabulary within which partial concepts are constructed, from the visual or other non-linguistic environment. The proposed presentation extends this argument to indefinite NP construal, which are defined as dependent terms. We will show how the construal of indefinite NPs is procedurally established over the course of a dialogue exchange, with the same range of forms of resolution and possibly with switch of speaker/hearer: dependency on a term already in context as in (1)-(3), dependency on a term to be subsequently locally constructed as in (4), and indexically (5):

(3) A: Will everyone in the competition need a...
B: a mattock? The ground is certainly very hard.

(4) A: A nurse interviews every patient
B: on which ward?
A: all of them. It is standard practice.

(5) A: A nice day at last.
B: Yes, isn't it.

Dynamic Syntax is uniquely well placed to model this phenomenon, as scope dependencies associated with quantifying expressions are induced on a step by step basis, these dependencies being defined as constraints on interpretation. There is thus a two-part construction process for quantifying terms: first, a process that induces the progressive construction of names, with scope dependency statements incrementally gathered together; second, an evaluation step in which the relationships of the constructed term to others within the overall construction are spelled out. At any point other than the final evaluation step, shift of roles is licensed, as in all cases this is made relative to a context having been constructed by either party, so no information is lost. Semantic construal of determiners is lexically defined and so allows variation across types, indefinites thus projecting an underspecified representation so that choice of scope dependency is determined by a free choice mechanism, analogous to pronouns. The apparent delay of projection of content in cases such as (5) is consistent with the incrementality requirement, being merely the anticipated combined effect of word-by-word processing and the update of partial specifications of content (analogous to expletive pronouns). The result is a characterisation of the flexibility of indefinites on a principled basis, enabling quantification construal, like all other aspects of natural-language structure and content, to be seen as grounded in mechanisms for coordinative interaction.

References

R. Cann, R. Kempson, and L. Marten. 2005. *The Dynamics of Language*. Elsevier, Oxford.

J. Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

E. Gregoromichelaki, R. Kempson, M. Purver, G. J. Mills, R. Cann, W. Meyer-Viol, and P. G. T. Healey. 2011. Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1):199–233.

C. Howes, M. Purver, P. G. T. Healey, Gregory J. Mills, and Eleni Gregoromichelaki. 2011. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311.

R. Kempson, W. Meyer-Viol, and D. Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.

R. Kempson. Fcmg. Syntax as the dynamics of language understanding. In K. Allan, editor, *Routledge Handbook of Linguistics*. Routledge.

M. Purver, R. Cann, and R. Kempson. 2006. Grammars as parsers: Meeting the dialogue challenge. *Research on Language and Computation*, 4(2-3):289–326.

Playing a Real-world Reference Game
using the Words-as-Classifiers Model of Reference Resolution

Casey Kennington
CITEC, Bielefeld University
ckennington@cit-ec.
uni-bielefeld.de

Soledad Lopez Gambino
CITEC, Bielefeld University
m.lopez_gambino@
uni-bielefeld.de

David Schlangen
CITEC, Bielefeld University
david.schlangen@
uni-bielefeld.de

Abstract

When referring to visually-present objects, an elementary site of language use, sometimes there isn't enough information to resolve the speaker's intended object. When this happens, more information needs to be elicited from the speaker. In this demo, we will show a simple system that uses the word-as-classifiers model to resolve referring expressions to objects, as well as a simple interaction manager that determines if there is enough information to fully resolve the reference–if not, more information is elicited from the speaker. The modules are implemented and distributed with InproTK.

1 Introduction

Reference to visually-present objects is a foundational language game. Among children's earliest communicative attempts are acts indicating objects for other people; for example, pointing to or displaying an object (Wittek and Tomasello, 2005) where the words of those references are *grounded* in the features of the objects being referred (Harnad, 1990). This setting of language use is situated dialogue where interlocutors can perceive each other, the objects in their shared space, and they can perceive each other's unfolding referring expressions (REs), often resolving the referred object before the RE is complete.

In this demo, we present a system that plays a similar language game: using the words-as-classifiers model of reference resolution (WAC$_{rr}$; explained below), we have a system that can resolve referring expressions (REs) incrementally to real-world objects with an additional component: an *interaction manager* (IM), that determines if more information should be elicited from the speaker.

In the following section we will describe the WAC$_{rr}$ model and how it fits into the system. That will be followed by a description of the interaction manager and the system implementation.

2 The Words-as-Classifiers Model of Reference Resolution

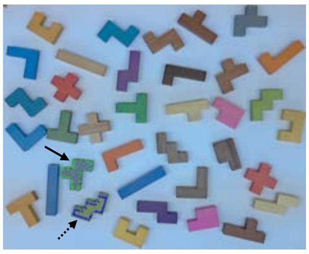


Figure 1: Example episode for where a referred *target* object is outlined in green, a *landmark* object (used to aid reference to the target) in blue; arrows added for presentation. An example RE for this would be: *the gray object on the bottom left above the green w*.

The basis of WAC$_{rr}$ is a model of *word meaning* as a function from visual features of an object to a judgment of how well that object "fits" a particular word.[1] The model can learn word meanings for picking out properties of single objects REs; e.g., *green* in *the green book* (Kennington et al., 2015) and picking out relations between two objects; e.g., *next to* (Kennington and Schlangen, 2015). These word meanings are learned from instances of language use.

These are then applied in the context of an actual reference. This application gives the desired result of a probability distribution over candidate objects, where the probability expresses the strength of belief in the object falling in the *extension* of the expression. We model two different types of composition, of what we call *simple ref-*

---

[1] This idea follows in spirit from Larsson (2013)

erences and *relational references*. These applications are compositional in the sense that the meanings of the more complex constructions are a function of those of their parts.

The meanings are represented as logistic regression classifiers. We train these classifiers using a corpus of REs coupled with representations of the scenes in which they were used (example in Figure 1) and an annotation of the referent of that RE. Meanings of relational words are trained in a similar fashion, except that they are presented a vector of features of a *pair* of objects, such as their euclidean distance.

During application, to get a distribution from a single word, we apply the word classifier to all candidate objects and normalise. To compose the evidence from individual words into a prediction for a 'simple' RE, we average the contributions of its constituent words. Relational REs are composed by combining two 'simple' REs via a learned classifier for a relation word. More details can be found in the two papers cited above in this section; they further show that the model is robust in reference resolution tasks despite noisy representations of scenes and speech (i.e., ASR).

3 The Interaction Manager

In a dialogue setting, making use of the distribution over objects requires an additional interaction manager which addresses certain cases in which the continuity of the game might be in jeopardy. This could happen if the user has not yet referred to any object (for example, when taking a long time to plan the RE) or if the speaker has already referred to an object but the information provided is not enough for the system to make a decision. Specifically, for this demo, the IM decides whether to select an object (i.e., the argmax of the distribution from WAC_{rr}) or if more information is needed.

4 Implementation

Figure 2 shows a schematic the overall system. The ASR (here, Kaldi[2]), WAC_{rr} and the IM have been implemented as modules in INPROTK (Baumann and Schlangen, 2012).[3] For the logistic regression classifiers in WAC_{rr}, we use the Apache Mahout Java library trained on a corpus of REs to objects in a scene.[4] We also have a module that can process a video feed of pentomino objects from a standard webcam (example in Figure 1) in real-time and provide the low-level features (e.g., RGB/HSV values, x,y coordinates, number of edges, etc.) of the scene to the WAC_{rr} module. The IM operates by reacting to lack of speech input from the ASR module. After a certain amount of time has elapsed and no voice activity has been detected, a timeout signal prompts the user to speak. Or, if the system has received an RE from the user but the information it contains is not enough to resolve the reference, after a certain amount of silence a simple clarification request is produced as a means of prompting the user to add further information.

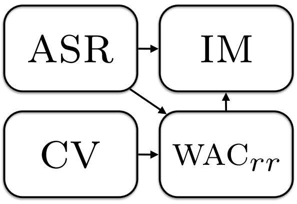


Figure 2: Scheme of the system: ASR and CV modules inform the WAC_{rr} module, which produces a distribution over objects; the IM module determines selection or elicitation.

Acknowledgements This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Timo Baumann and David Schlangen. 2012. The InproTK 2012 Release. In *Proceedings of NAACL-HLT*.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D*, 42:335–346.

Casey Kennington and David Schlangen. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of ACL*, Beijing, China. Association for Computational Linguistics.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of IWCS*. Association for Computational Linguistics.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

Angelika Wittek and Michael Tomasello. 2005. Young children's sensitivity to listener knowledge and perceptual context in choosing referring expressions. *Applied Psycholinguistics*, 26(04):541–558.

---

[2] `http://kaldi.sourceforge.net`

[3] `https://bitbucket.org/inpro/inprotk`

[4] `http://mahout.apache.org/`

The State of the Art in Dealing with User Answers

Staffan Larsson
Talkamatic AB and University of Gothenburg
Gothenburg, Sweden
`staffan@talkamatic.se`, `sl@ling.gu.se`

1 Introduction

This paper takes a look at how dialogue systems from three major players (Apple, Google, and Microsoft) on the market actually deal with some conversational behaviours frequently encountered in human-human dialogue[1]. The work presented here builds on the "Trindi Tick-list" (Bos et al., 1999).

2 Over-answering

What happens if the user gives more information than the system asked for? To keep it simple, we assume that the extra information given is still relevant to the task.

U: Make a call
S: OK, who do you want to call?
U: John
S: What number?
U: Mobile
S: OK, placing the call.

The dialogue above does feel a bit tedious; why not speed it up a little by giving more than one bit of information at a time?

U: Make a call
S: OK, who do you want to call?
U: John's mobile
S: OK, placing the call.

Note that S, when responding to U's question about who to call, also provides the desired kind of number (mobile).

At least two of the systems we tested do not seem to handle over-answering very well. Here's how one of them (Google Now) responds:

U: Make a call
S: OK, who do you want to call?
U: Fred's mobile
S: Try repeating the contact name or say Cancel

On the positive side, the two other systems we tested (Siri and Cortana) deal well with over-answering, and behave as expected.

3 Other-answering

We will now look at the related phenomena of other-answering. What if S had been mentally focusing on the number type rather than the person to call, and blurted out "mobile" in response to U's questions about who to call? A graceful continuation by the system would be to repeat the question about who to call.

U: Make a call
S: OK, who do you want to call?
U: Mobile
S: OK, mobile. Who do you want to call?

It turns out that Google Now has trouble with getting an answer to another question than the one they asked.

S: Who do you want to call?
U: Mobile
S: Say the name of the contact you want to call, or say Cancel

And here is the Cortana response:

S: Who would you like to call?
U: Mobile
S: I couldn't find that contact.

Basically, it seems that all these systems rely heavily on the assumption that the user will provide the information asked for. Whatever occurs after the question will be interpreted as an answer to it, even if it is not an appropriate answer (a name

---

[1] Some of the results presented here have previously been published in blog post format on the Talkamatic blog, `http://www.talkamatic.se/category/blog/`.

	Google Now	Siri	Cortana
Over-answering	misunderstood	+	+
Other-answering	misunderstood	+	ignored
Embedded revision	ignored	halts	halts

Table 1: Dialogue behaviours related to answering questions in three dialogue systems

in the phonebook), and even if it is a perfectly appropriate answer to a question that the system is just about to ask (about number type).

However, one system (Siri) does deal with other-answering in the way we outlined above. As an example, we use the email domain[2].

U: Send an email
S: To whom shall I send it?
U: It's about the trip
S: OK, I updated the subject [showing "Trip" as subject]
S: Ready to send it. To whom shall I send it?

As we can see, Siri also explicitly acknowledges understanding and integrating the "other" information.

4 Answer revision

By *answer revision*, we mean cases where a resolved question is given a new answer intended to replace the previous one. The desired behaviour from the system in such cases is that the old answer should be retracted and replaced by the new answer. Here is an example:

U: Call Jane
S: OK, Jane. What number?
U: No, Fred
S: OK, Fred. What number?
U: Mobile
S: Calling Freds mobile.

Siri ignores bare revisions altogether:

U: Call Jane
S: Which phone number for Jane?
U: Fred
S: Which phone number for Jane?

Embedded corrections starting with "no" result in the dialogue grinding to a halt:

U: Call Jane
S: Which phone number for Jane?
U: No, Fred
S: OK. (nothing happens)

Our second system (Google Now) does not seem to handle any kind of answer revision, regardless of form:

U: Call Jane
S: Home or mobile? [showing "Jane"]
U: Fred / No, Fred
S: Home or mobile? [still showing "Jane"]
U: mobile
S: Calling Jane's mobile.

Cortana ignores bare revisions, similar to Google Now. For embedded revisions, Cortana seems to misunderstand and get the "no" but not revised answer, thus ending up asking for a name that the user has already provided:

U: Call John
S: What number? Mobile or work?
U: No, Peter
S: Sure, who do you want to call?

5 Conclusions and future work

Our results are summarized in Table 1. Google Now does not do very well, which indicates that perhaps it is not intended as a full dialogue system. Cortana is also not very successful. Siri does quite well, but there is still room for improvement. As existing systems are improved and new systems appear on the market, investigations such as the one presented here need to be continually revised.

References

Johan Bos, Staffan Larsson, I Lewin, C Matheson, and D Milward. 1999. Survey of existing interactive systems. Technical Report D1.3, TRINDI (Task Oriented Instructional Dialogue) project.

---

[2] The reason for using the email domain is that in the phone domain, if the user does not respond to the question about who to call, Siri assumes that the person to call is the user her/himself. We leave a discussion of such behaviours for another paper.

Polish Event-Linking Devices of *przed* ‘before’ cluster in conversational data – implications for Contrastive Analysis

Barbara Lewandowska-Tomaszczyk
University of Lodz, Institute of English Studies, Pomorska 171/173, 90-236 Lodz, Poland
blt@uni.lodz.pl

Abstract

The poster is part of a larger project on *Event-Linking Devices* (ELDes), and aims to look into one category of Event-Linking phenomena, the concepts of *przed* ‘before’ cluster in Polish, and their semantic functions of precedence/succession, priority/posteriority, and others in discourse (Lewandowska-Tomaszczyk et al., 2015). The study is based on empirical data derived from referential corpora of Polish (nkjp.pl), as well as, to contrast it with English, from the translational (parallel) English-to-Polish and Polish-to-English corpora available at http://pelcra.clarin-pl.eu/. The focal research questions refer to the uncovering of paths which account for coherent linking of events and their parts in the case of Pol. *przed* ‘before’ and its adjacent cluster members, with instances of the cognitive *afterness* called forth where relevant.

1 Introduction

Before and *after* are, as suggested by Östen Dahl (2013), *time-creating* conceptual areas. It is argued that *przed* is associated with a *scale of senses* such as the most salient ones including the cognitively basic *object-linking spatial* sense, extended to cover *temporal*, *sequential*, *contrastive (confrontational)*, and *conditional* interpretations, which, by extension, involve either *event chains* (coordination) or *event hierarchies* (subordination).

2 Research methodology

The research methods used are both quantitative, i.e., considering the frequencies of use of particular forms, as well as qualitative, i.e., involving the cognitive frame-based linguistic and discourse perspectives. The study presents an analysis of Polish corpus data of the *przed cluster* for non-annotated discourse relations, with English translational equivalents of ambiguous connectives (Cartoni et al., 2013), and English parallel corpus data and their functional interpretation.

3 Frames and Reframing

The linguistic *przed/before* clusters activate an original *spatial frame* in which **physical objects** are positioned in terms of primary versus secondary focal (spatial) positions (one object positioned before another object) *przed nim* ‘before (in front of) him’, *przed telewizorem* ‘before (in front of) TV’. A range of *before*-senses is extended by *re-framing* the original spatial relations (Lakoff and Johnson, 1980) into the temporal *before* ones *przed chwilą* ‘a second ago’, *przed deszczem* ‘before (it started) raining’. Further extensions, in complex constructions, cause the mapping of other Target Domains such as *succession/consequence* Eng. *pride goeth <u>before</u> destruction*, *primacy/priority* and *condition*, *concession*, *causality, nie mam <u>przed</u> tobą tajemnic* ‘I have no secrets <u>before</u> you (positive confrontation)’, *<u>przed</u> niczym się nie cofnę* ‘I will not go back before anything (negative confrontation)’, *causality-effect* Eng. *Put the cart <u>before</u> the horse.* The relative frequencies of the senses cover spatial meanings - 30 % of all data examined, temporal - 62 %, confrontational, conditional, priority and others - 8%.

4 Event Linking Devices

Events linked by *przed* are expressed either by a Nominalized gerund or Verbal noun construction *przed zakończeniem/przyjazdem* ‘before finishing/arrival’, or, in case the event is expressed by a clause, by *przed* taking up a complex form: *przed tym (zanim)* or *(przedtem) (zanim)* lit. ‘before that (by the time before)’. The range of senses covered in conversational materials include temporal (most frequent) and, in descending order: contrastive (negative confrontation, invariably introduced by *threat/fear* forms), (temporal) conditional (most frequently introduced by the negative *zanim nie* lit.’ before not’ (not until/unless), and sequential meanings.

jej trwoga	**przed tym**	, co za chwilę nastąpi

‘her fear **before that**, what will happen in a moment’ (negative confrontation, challenge)

ja cie nie wpuszczę do mojego mieszkania	**Zanim**	naprawdę będzie. jak będzie skończone to drugie

‘I will not let you in to my flat until/before (unless) the other one is finished’ (condition)

4.1 Frequencies

All categories of *przed* were first searched in the whole Polish (balanced) corpus, comprising 250 m units with *przed* constructions identified in 253,119 cases. In spoken data the frequency reached 1,494, while in *conversational* materials (ca. 1,5 m), it was 778. For the clause-initial phrase *przed tym,(zanim)*, the frequencies in the whole corpus did not exceed 80 and in the conversational materials the occurrence was below 10 for each. *Przedtem* and *zanim* and their combination have the highest frequency of occurrence (7,882/25,892/74) in all materials with frequencies not exceeding 200 for each in conversational data:

proszę pana ja nie widzę tu w ogóle jakiegokolwiek. a dlaczego	**przedtem**	on nie reklamował ?

‘Sir, I can’t see any here, and why didn’t he intervene before?’

Nie mogę mówić o tym,	**zanim nie**	uzyskam pewności

‘I can’t talk about it unless (until) I am fully certain’

4.2 Other *przed* cluster members

Forms considered discursively synonymous with the temporal sense of *przed* are *najpierw* lit. ‘first’, which take a quantitative preference over *przed* in the total corpus (39,869) and in the conversational data reaches 519 occurrences, *poprzedzając* ‘preceding/preceded’ (42) and a number of others. A combinatorial form which deserves particular attention in Polish is a pair involving the opposite items *przedtem – potem* ‘before-after’, which however function as practical synonyms in the context of the temporal *przedtem (najpierw)* and *potem jak* (but not *przedtem jak*) combinations in the sense of *succession* (*sequence*, possibly *causal*), revealing the *time* to *succession/cause-consequence* frame-shifting:

Przedtem (Najpierw) zjadłam lody, a potem zachorowałam

(lit. Before)/First I ate ice-cream and then I got sick

Potem jak zjadłam lody zachorowałam.

‘After (lit. after how (when)) I ate ice-cream, I got sick’ (more frequent in Conversational Data).

5 Further research

Further study is aimed to compare *przed* with *po* ‘after’ clusters (i.a., with respect to links to spatial and other conceptual domains) as well as to contrast other Polish and English markers, particularly those containing elements of negativity (Lewandowska-Tomaszczyk 2004) and used in emotional contexts, and propose a typology of ELDes in Polish and English, in spoken and written modes. Implications to cross—linguistic study of such phenomena will be presented in order to provide some more explicit ELDes *annotation clues* in the case of complex discourse-related meaning phenomena and *cluster equivalence* in languages.

Acknowledgment

The study is carried out within the COST Action IS1312 *Structuring Discourse in Multilingual Europe* (TextLink).

References

Bruno Cartoni, Sandrine Zufferey and Thomas Meyer. 2013. Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique. In Stefanie Dipper, Heike Zinsmeister, and Bonnie Webber (eds). *Discourse and Dialogue. Beyond semantics: the challenges of annotating pragmatic and discourse phenomena*, volume 4, No 2. 65-86.

Östen Dahl. 2013. How Telicity Creates Time. *Journal of Slavic Linguistics* 21(1). 45-76.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago.

Barbara Lewandowska-Tomaszczyk, Piotr Pęzik, Paul A. Wilson and Jerzy Tomaszczyk. 2015. *DRDs in a contrastive perspective: a corpus-based cognitive study*. Louvain-La-Neuve: 1st conference COST Action IS1312 *Structuring Discourse in Multilingual Europe* (TextLink).

Barbara Lewandowska-Tomaszczyk. (2004) Conceptual blending and discourse functions. *Research in Language* 2. 33-47.

Developing Spoken Dialogue Systems with the OpenDial toolkit

Pierre Lison
Language Technology Group
University of Oslo (Norway)
plison@ifi.uio.no

Casey Kennington
Dialogue Systems Group, CITEC
Bielefeld University (Germany)
ckennington@cit-ec.uni-bielefeld.de

Abstract

We present OpenDial[1], an open-source toolkit for building and evaluating dialogue systems. The toolkit is centered on a dialogue state expressed as a Bayesian network and acting as a shared memory for the system modules. The domain models are specified via *probabilistic rules* encoded in a simple XML format. The toolkit has been deployed in several applications domains such as human–robot interaction and in-car driver assistants.

1 Introduction

Software frameworks for dialogue systems are often grouped in two categories. Symbolic frameworks rely on finite-state or logical methods to represent and update the dialogue state. While they provide fine-grained control over the dialogue flow, these approaches often assume that the state is fully observable and are thus poor at handling errors and uncertainty. Statistical frameworks, on the other hand, can automatically optimise the dialogue behaviour from data. However, they generally require large amounts of training data, making them difficult to apply in data-scarce domains.

The OpenDial toolkit adopts a hybrid approach that combines the benefits of logical and statistical methods into a single framework. The toolkit relies on *probabilistic rules* to represent the internal models of the domain in a compact and human-readable format. Crucially, the probabilistic rules may include unknown parameters that can be automatically estimated from data using supervised or reinforcement learning.

2 Architecture

The OpenDial toolkit relies on an information-state architecture in which all components work together on a shared memory that represents the current dialogue state (see Fig. 1). This state is encoded as a Bayesian network, where each variable captures a particular aspect of the interaction.

The toolkit itself is domain-independent. To apply it to a given dialogue domain, the system developer provides an XML-encoded domain specification containing (1) the initial dialogue state and (2) a collection of probabilistic rules for the domain. The probabilistic rules are expressed as *if...then...else* constructions mapping logical conditions to probabilistic effects. The rule conditions are logical formulae using the standard operators and relations from predicate logic. Each condition is associated with a distribution over mutually exclusive effects, where each effect corresponds to an assignment of values to some state variable(s). The parameters of these distributions may be initially unknown, allowing them to be estimated from data via Bayesian learning.

At runtime, the probabilistic rules are instantiated in the Bayesian network representing the dialogue state. Standard algorithms for probabilistic inference are then employed to update the dialogue state and select new system actions. Lison (2015) provides more details about the formalisation of probabilistic rules, their instantiation and the estimation of their unknown parameters.

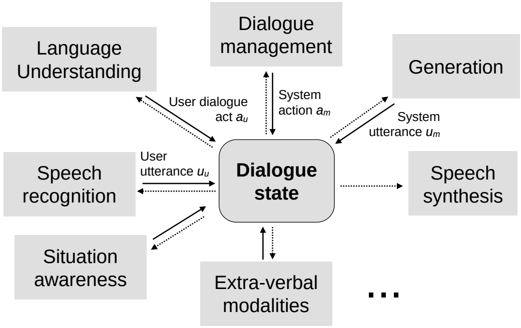
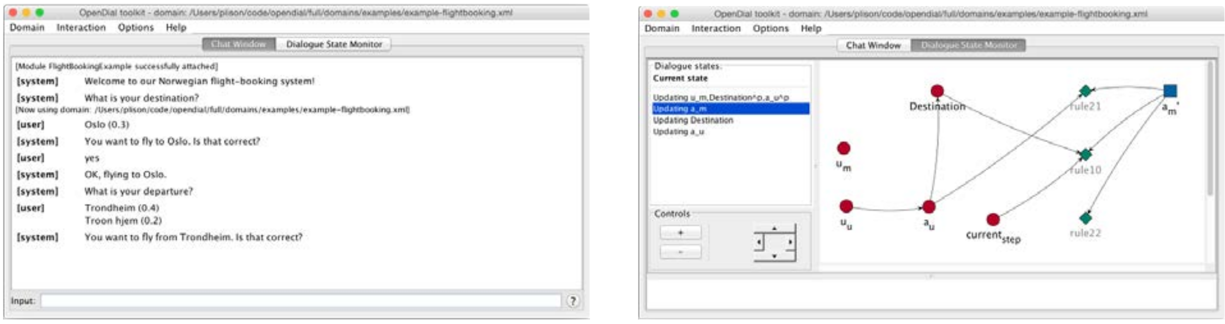


Figure 1: Schematic architecture for the toolkit.

---

[1] http://opendial-toolkit.net.



```
[Module FlightBookingExample successfully attached]
[system]      Welcome to our Norwegian flight-booking system!
[system]      What is your destination?
[Now using domain: /Users/plison/code/opendial/full/domains/examples/example-flightbooking.xml]
[user]        Oslo (0.3)
[system]      You want to fly to Oslo. Is that correct?
[user]        yes
[system]      OK, flying to Oslo.
[system]      What is your departure?
[user]        Trondheim (0.4)
              Troon hjem (0.2)
[system]      You want to fly from Trondheim. Is that correct?
```

(a) Chat window

(b) Dialogue state monitor

Figure 2: Graphical user interface for the toolkit.

3 Implementation and Applications

The toolkit is implemented in Java and released under an open-source license. The software comes with a graphical user interface allowing system developers to run a given dialogue domain and test its behaviour in an interactive manner (see Fig. 2). A collection of plugins extends the toolkit with external modules for e.g. speech recognition and synthesis or dependency parsing.

The toolkit has been deployed in several application domains, either as an end-to-end system or as a sub-component in a larger software architecture. One application domain is human–robot interaction. Lison (2015) illustrates the use of OpenDial in a domain where a Nao robot is instructed to navigate through a simple environment, fetch an object and bring it to a landmark. The probabilistic rules for the domain relied on parameters estimated from Wizard-of-Oz data. Similarly, Kennington et al. (2014a) describe how the toolkit was used as dialogue manager in a twenty-questions game between a robot and up to three human participants. The parameters of the dialogue policy were also estimated using Wizard-of-Oz data. This domain was not only multi-participant, but also multi-modal: the toolkit tracked the state of each individual participant's speech, dialogue act, attention (e.g. towards the robot or towards another participant), and participation state (e.g. passive or active) and provided decisions on what to say and where to direct the robot's attention.

The toolkit has also been deployed as dialogue manager in an in-car dialogue scenario (Kennington et al., 2014b). The system objective was to deliver upcoming calendar entries to the driver via speech. The toolkit was employed to track the state of the dialogue over time (using information from the driving simulator to make the system "situationally aware") and decide when the system's speech should be interrupted, for example when a dangerous driving situation (e.g. a lane change) was detected. This system was later enhanced to allow the driver to signal via speech or a head nod that the interrupted speech should resume.

4 Conclusion

The OpenDial toolkit rests on a hybrid framework combining ideas from both logical and probabilistic approaches to dialogue modelling. The dialogue state is represented as a Bayesian network and the domain models are specified using probabilistic rules. Unknown parameters can be estimated from dialogue data via Bayesian learning.

The toolkit is in our view particularly well-suited to handle dialogue domains that exhibits both a complex state-action space and partially observable environments. Due its hybrid modelling approach, the toolkit is able to capture such dialogue domains in a relatively small set of rules and associated parameters, allowing them to be tuned from modest amounts of training data, which is a critical requirement in many applications.

References

C. Kennington, K. Funakoshi, Y. Takahashi, and M. Nakano. 2014a. Probabilistic multiparty dialogue management for a game master robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 200–201.

C. Kennington, S. Kousidis, T. Baumann, H. Buschmeier, S. Kopp, and D. Schlangen. 2014b. Better Driving and Recall When In-car Information Presentation Uses Situationally-Aware Incremental Speech Output Generation. In *Proceedings of Automotive'UI*, Seattle, USA.

P. Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232 – 255.

CCG for Discourse

Sumiyo Nishiguchi
School of Management, Tokyo University of Science
500 Shimokiyoku, Kuki-city, Saitama, 346-8512, Japan
`nishiguchi@rs.tus.ac.jp`

Abstract

Question and answer congruence has been considered to be a discourse unit (Groenendijk and Stokhof 1984, Ginzburg and Sag 2001). I propose a new framework for question-answer pairs and focused sentences in Combinatory Categorial Grammar (CCG) (Steedman 2000). In CCG, questions and focused sentences have been assigned the categories of S (Jäger 2005, Barker and chieh Shan 2006), while questions are sets of possible answers semantically (Hamblin 1973). Pragmatically, focus induces a set of alternatives (Rooth 1992). I claim that interrogatives and focused sentences should be functions from a sentence to another sentence in view of their semantics. Such novel categories enable combining with the following sentence in a discourse by functional application. Thereby, Japanese sentence-final particles such as a question marker *ka* are category S\(S/S), and *yo* and *no* are polarity focus operators (Höhle 1992).

1 Modality in CCG and TLG

In CCG, questions, focused sentences and exclamatives have been considered to be of the category S, or a sentence. Steedman (2000) specifies features for focused sentences and uses prosodically annotated categories. Barker and chieh Shan (2006), in multi-modal TLG, introduces a modality $\circ ?$ to term questions. The category of a question sentence is $_?$S and the lexical category of *what* is (NP\$_?$S)/(NP\S) which combines with a predicate and returns a function from NP to a question sentence. Jäger (2005) terms questions as the category *q*, and *wh*-phrases $q/(np\uparrow s)$, a function from a predicate to a question. In Hockenmaier and Steedman (2007), S carries a feature— declaratives (S[dcl]), *wh*-questions (S[wq]), *yes-no* questions (S[q]), or fragments (S[frg]). Even though the proposed modalized sentence categories are useful for controlling combinatorics, such modality is not really necessary if syntax-semantics correspondence is more strictly pursued.

2 Proposal: Higher Order for Questions and Polarity Focus

Syntactic categories should reflect the semantic content of questions and focused sentences.

2.1 Semantics of Questions and Focus

Semantically speaking, the denotation of a question or a focused sentence is assumed to be a set of propositions. For example, the interpretation of *Did you see Alice* is a set of possible answers in a given context (Hamblin 1973, Kartunnen 1977):

(1) $[\![Did\_you\_see\_Alice?]\!]$ = {you saw Alice, you did not see Alice}

Since a proposition is a set of possible worlds which is of type <s, t>, the set of possible answers is a set of sets of possible worlds, <st, t> . Focus induces sets of alternative propositions (Rooth 1992). In (2a), the alternative answers along with the real answer form a set of contextually possible answers called focus semantics value (“f”) without truth-conditional contribution.

(2) a. A: Where did you go on weekend?

B: I went to the BEACH.

b. $[\![I\_went\_to\_the\_BEACH]\!]^f$

= {I went shopping, I went hiking, I stayed home,...}

2.2 New Lexical Category for Questions and Focus

The semantic type of questions and focused sentences <st, t> more straightforwardly correspond to type S/S rather than S_Q or S_{foc} even though there is no syntactic composition of two sentences. Therefore, I propose the following lexical entries.

(3) a. A polar question: S/S: {p, ¬p}

b. A focused sentence: S/S: {p, q, r,...}

Such novel categories can handle discourse:

$$\dfrac{\dfrac{\dfrac{who}{(S/S)/(NP\backslash S):\lambda f_{<et>},p_{<t>}.\pi(p)}\text{Lex}\qquad \dfrac{came}{NP\backslash S:\lambda x.came'(x)}\text{Lex}}{S/S:\lambda p.\pi(p)}>\qquad \dfrac{\dfrac{Mary}{NP:m}\text{Lex}\qquad \dfrac{did}{NP\backslash S:\lambda x.f(x)}\text{Lex}}{S:f(m)}<}{S:\pi(f(m))}>$$

The syntactic category of the question *who came* is S/S which combines with the answer by means of inter-sentential functional application.

(4) Functional Application

A/B: f, B: a ⇒ A: f(a) (>)

A: a, A\B: f ⇒ B: f(a) (<)

2.3 Fragmental Answers as Propositions

If we consider questions as sets of propositions, how would the questions combine with fragmental answers in the forms of NPs, VPs, or PPs that are not full Ss? From pragmatic viewpoint, Stainton (2004) considers assertions of non-propositional fragments as type <t>.

(5) A: What did you eat?

B: Apples.

Although "Apples." is a noun phrase whose semantic type is <et, t>, its pragmatic contribution is the same as "I ate apples" of type <t>, which is the form before ellipsis. In the present analysis, inter-sentential functional application requires the response to be category S. As the pragmatic contribution of fragmental answers are of S and not NP, PP or VP, semantic type-raising of fragmental answers to the category S makes functional application possible between questions and fragmental answers.

2.4 Question-question Congruence

Sometimes questions are replied with another question in cases of presupposition failure. In (6), the speaker's presupposition that the proper name *Alice* has reference is not shared by the hearer. Functional application cannot be applied because the second question of type S/S cannot be an argument of the first question of type S/S. Instead of functional application, functional composition is necessary (Curry and Feys 1958, Steedman 2000).

(6) a. Forward Composition (>B)

A/B B/C $\rightarrow_B$ A/C

b. Backward Composition (<B)

A\B B\C $\rightarrow_B$ A\C

(7) A: Did you see Alice?

B: Who is Alice?

$$\dfrac{\dfrac{\dfrac{Did}{(S/S)/S:\lambda p_{<t>},p.\pi_{<t,t>}(p)}\text{Lex}\qquad \dfrac{\dfrac{you}{NP:h}\text{Lex}\qquad \dfrac{\dfrac{see}{(NP\backslash S)/NP:\lambda x,y.see'(x)(y)}\text{Lex}\quad \dfrac{Alice}{NP:a}\text{Lex}}{NP\backslash S:\lambda y.see'(a)(y)}>}{S:see'(a)(h)}<}{S/S:\lambda p.\pi(see'(a)(h))}>\qquad \dfrac{\dfrac{Who}{(S/S)/(NP\backslash S):\lambda f_{<et>},p.\rho(p)}\text{Lex}\qquad \dfrac{\dfrac{is}{(NP\backslash S)/NP:\lambda x,y.be'(x)(y)}\text{Lex}\quad \dfrac{Alice}{NP:a}\text{Lex}}{NP\backslash S:\lambda x.be'(a)(x)}>}{S/S:\lambda p.\rho(p)}>}{S/S:\lambda f.\lambda p.\pi(see'(a)(h))((\lambda p.\rho(p))(f))}>\text{B}$$

References

Barker, Chris, and Chung chieh Shan. 2006. Types as graphs: Continuations in type logical grammar. *Journal of Logic, Language and Information* 15:331–370.

Curry, Haskell B., and Robert Feys. 1958. *Combinatory logic i.* Amsterdam: North Holland.

Ginzburg, Jonathan, and Ivan Sag. 2001. *Interrogative investigations.* Stanford: CSLI Publications.

Groenendijk, Jeroen, and Martin Stokhof. 1984. Studies on the semantics ofquestions and the pragmatics of answers. Doctoral Dissertation, University of Amsterdam.

Hamblin, C. L. 1973. Questions in montague grammar. *Foundations of Language* 10:41–53.

Hockenmaier, Julia, and Mark Steedman. 2007. Ccgbank: A corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics* 33:355–396.

Höhle, Tilman N. 1992. Über verum fokus in deutschen. *Linguistische Berichte* 4:112–141.

Jäger, Gerhard. 2005. *Anaphora and type logical grammar.* Dordrecht: Springer.

Kartunnen, Lauri. 1977. Syntax and semantics of questions. *Linguistics and Philosophy* 1:1–44.

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

Stainton, Robert J. 2004. The pragmatics of non-sentences. In *The handbook of pragmatics*, 266–287. London: Blackwell.

Steedman, Mark. 2000. *The syntactic process.* Cambridge, Mass.: MIT Press.

Experimenting with grounding strategies in dialogue

Volha Petukhova[1], **Harry Bunt**[2], **Andrei Malchanau**[1], **Ramkumar Aruchamy**[1]
[1] Saarland University, Spoken Language Systems, Germany; [2] Tilburg University, Netherlands
{v.petukhova,andrei.malchanau,ramkumar.aruchamy}@lsv.uni-saarland.de harry.bunt@uvt.nl

Abstract

This paper discusses empirically grounded strategies for the generation of feedback acts by a dialogue system in a way that supports a natural communication style and therefore leads to higher user acceptance. User evaluation of an implemented prototype system shows that an appropriate strategy can be generated by rules that are based on an analysis of human-human dialogue behaviour for a given this task and domain.

1 Introduction

While conversational speech-based applications have recently begun penetrating the mass market, commercial dialogue systems are still limited to a rather restricted communication behaviour modelled on information providing tasks. Some systems developed for research purposes allow for more natural conversations, but they are often limited to a narrow domain with manually crafted domain models and pre-baked dialogue strategies. Alternatively, dialogue strategies can be adapted through reinforcement learning, but this requires large amounts of training data, while offering only a limited range of dialogue actions.

In this paper we show how a relatively small amount of 'Wizard-of-Oz'(WoZ) data and focused analysis of the phenomena related to grounding can help to design various strategies and communicative styles in order for a dialogue system to exhibit behaviour that is more natural to its users.

2 Observed grounding behaviour

To simulate user's information-seeking and system's information-providing behaviour we designed a set of quiz games. Data has been collected in a WoZ setting with the Wizard holding the facts about a famous person's life, and a player guessing his/her identity by asking questions of various types. 338 dialogues were collected (16 hours comprising about 6.000 speaking turns, 18 turns per dialogue), transcribed and annotated with ISO 24617-2 dialogue acts.[1]

For an interactive system it is important to know that its contributions are understood and accepted (i.e. grounded) by the user. In our quiz scenario, if the answer is understood and accepted by the player, he continues with his next question. However, we do not just observe question-answer pairs. Players very often signal their understanding and acceptance of the previous system utterance by repeating or rephrasing (part of) it, known as 'implicit verification', or accepting answers with inarticulate positive feedback like 'Okay', 'mm-mhm', 'yeah', 'right', etc. This allows the user to verify the correctness of the system's recognition of the preceding utterance, and gives the user the possibility to correct mistakes on the fly (allo-feedback). In case of positive feedback from a player, the Wizard often explicitly acknowledges it, and in case of negative feedback always reacts to it.

We analysed the data for the occurrence of sequences of Questions, Answers, positive/negative Auto- and AlloFeedback acts. Table 1 presents the frequencies of the patterns that were observed. These patterns were used to construct a decision tree for feedback generation, weighting possible transitions from one state to another. It may be observed that the simple Question-Answer sequence is the most frequent pattern, however explicit positive Auto-Feedback occurs quite often.

A dialogue system that provides positive auto-feedback after every user contribution would exhibit a style of communicative behaviour is unnatural and even annoying. It is therefore interesting

---

[1] For the ISO 24617-2 dialogue act annotation standard see Bunt et al., 2012; for details on the data collection and the annotation see Petukhova et al., 2014.

Observed sequence	Frequency (in %)
P:Question1 - W:Answer - P:Question2	47.1
P:Question1 - W:pos. AutoFeedback - W:Answer – P:Question2	28.6
P:Question1 - W:neg. AutoFeedback(execution: answer not found) - P:pos. AutoFeedback - P:Question2	7.6
P:Question1 - W:neg. AutoFeedback - P:Repeat/rephrase Question - W:pos. AutoFeedback - W:Answer - P:Question2	6.3
P:Question1 - W:pos. AutoFeedback - W:Answer - P:pos. Allo/AutoFeedback - P:Question2	4.9
P:Question1 - W:neg. AutoFeedback(execution: answer not found) - P:Question2	2.8
P:Question1 - W:neg. AutoFeedback - P:Repeat/rephrase Question - W:Answer - P:Question2	2.7

Table 1: Observed sequences of player-system acts ranked according to relative frequencies.

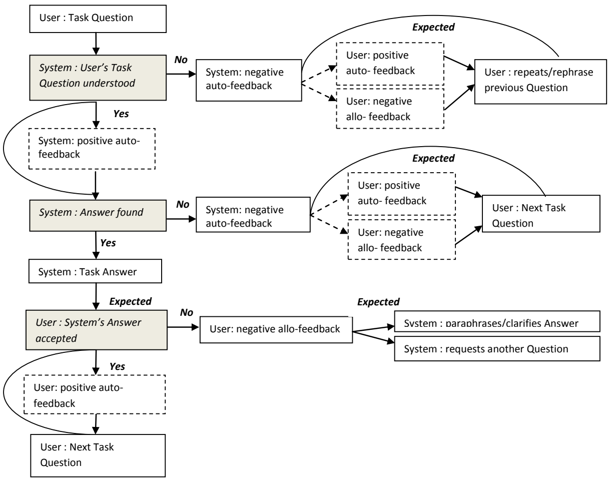

User : Task Question
System : User's Task Question understood
No
System: negative auto-feedback
Expected
User: positive auto- feedback
User: negative allo- feedback
User : repeats/rephrase previous Question
Yes
System: positive auto-feedback
System : Answer found
No
System: negative auto-feedback
Expected
User: positive auto- feedback
User: negative allo- feedback
User : Next Task Question
Yes
System : Task Answer
Expected
User : System's Answer accepted
No
User: negative allo-feedback
Expected
System : paraphrases/clarifies Answer
System : requests another Question
Yes
User: positive auto-feedback
User : Next Task Question

Figure 1: Decision tree for the generation of dialogue acts by the system. Dashed boxes present optional actions; gray boxes represent actual or expected processing states.)

to consider strategies where positive feedback is generated *regularly* when the dialogue reaches a crucially important state, and only *occasionally* in other situations, e.g. when it is not vital for task performance, and *regularly*, e.g. when the dialogue reaches a crucially important state. For our scenario and dialogue setting we designed a decision tree that incorporates the observations and analysis of our data for the generation of various types of feedback, see Figure 1.

To evaluate to what extent users feel that grounding strategies modelled in such a way lead to natural and flexible interactive behaviour, three question-answering system prototypes were implemented using the NPCEditor tool[2], extening the dialogue management strategies defined in the NPCEditor in order for the system to show more complex interactive behaviour beyond question-answering, by adding more dialogue manager states and a wider variety of positive and negative auto- and allo-feedback act types.

For the evaluation we investigated user satisfaction using a questionnaire filled in after interacting with the system. A within-subject evaluation was performed with 6 users who played a game using three different system prototypes: (1) minimal query - response (MQR) setting; (2) system always generating explicit auto-feedback to player's query (AEFR); and (3) system generating explicit feedback according to the decision tree shown in Figure 1 (DEFR).

We tested user satisfaction by asking subjects to rate their level of agreement on the following parameters: (i) learnability (the ease to use the system, e.g. rules well explained); (ii) ability to get the requested information; (iii) correctness of answers; (iv) frequency and type of system feedback; (v) speed of responses; (vi) naturalness of the interaction and (vii) overall attitude, e.g. likability and engagement. For each parameter we obtained the agreement scores. Responses for each question were summed up and divided by the number of participants to calculate the level of agreement in terms of average Likert scores.[3] The results show that players in general appreciate explicit feedback, and when the system generated feedback acts according to the decision tree it received the highest score on all criteria without exception: MQR was rated 3.4 on 5-point Likert scale; AEFR - 3.6; and DEFR - 4.5.

This exploratory study left some unexplored and/or not implemented options. For instance, the behaviour in other dimensions than the feedback dimensions such as Turn-, Time-, Own- and Partner Communication Management, and Discourse Structuring deserves attention; findings there may well lead to more interesting and flexible behaviour on the part of the system.

References

Bunt, H. et al. 2012 ISO 24617-2: A semantically-based standard for dialogue annotation. In: *Proceedings of LREC 2012*, Istanbul, Turkey

Petukhova, V. et al. 2014 The DBOX corpus collection of spoken human-human and human-machine dialogues. In: *Proceedings of LREC 2014,* Reykjavik, Iceland

---

[2] `https://confluence.ict.usc.edu/display/VHTK/NPCEditor`

[3] Statistical tests were not performed, since the number of raters (6) was too low to draw statistically significant conclusions. Our goal was rather to obtain first impressions of the acceptance or rejection of different grounding strategies by human players.

Self Awareness for Better Common Ground

Ondřej Plátek and **Filip Jurčíček**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{oplatek,jurcicek}@ufal.mff.cuni.cz

Abstract

Having set up several dialogue systems for multiple domains we realized that we need to introduce each system thoroughly to the users in order to avoid misunderstanding. Even with well-described systems, users still tend to request out-of-domain information and are often confused by the system response. In our ongoing work, we try to address these issues by allowing our conversational agent to speak about its abilities. Our agent is able to simply describe what it understands and why it decided to perform one of its summary actions. In this short paper, we present our current system architecture.

1 Introduction

Searching for a common ground in spoken conversation is a well-studied process known as *grounding* (Traum, 1994). We argue that out-of-domain questions are a-type of misunderstanding where the user incorrectly assumes a capability of the system. As a result, we think that dialogue systems should inform the user better about what their communication possibilities are.

Research in dialogue systems has focused so far on improving goal-oriented narrow domain systems. Such systems unfortunately communicate only about their domain and users easily request out of domain information. The systems often cannot help the user if the user is lost in the conversation because it is hard to come up with correct response in out of domain context.

In our effort to develop conversational systems which try to learn domain knowledge from users, we realized that the first thing that users are trying to teach the system is to communicate better. One of our goals is to develop a policy which allows the system to confirm if the misunderstanding occurred, and identify its source. We attribute the misunderstanding to one of the system components language understanding (LU), dialogue management (DM) or natural language generation (NLG). Therefore we would like to confirm if the system understood the user badly, chose surprising action or used wrong words.

The first step is to present the user which actions the system is able to take and how it represents the dialogue history. In our system, we explicitly store information about the system, so the agent can communicate about its abilities. On the other hand, we do not attempt to explain to our users which facts the system is able to communicate about, and which facts are not present in its knowledge base. Consider example conversation between user **U** and system **S** to see how information about action *ask*, *slot date*, and *LU* component can help the user understand the system behaviour.

U: *Can you make a reservation for tonight evening at 20:00?*
S: *What movie do you want to see?*
U: *edge of tomorrow*
S: *Do you want to change the reservation for tomorrow?*
U: *No! What?*
S: *I asked you about the date because you change it by saying tomorrow. Do I understand you incorrectly?*
U: *Yes, I told you the name of the movie which is Edge of Tomorrow.*

It is hard for the user to imagine what she or

This research was partly funded by the Ministry of Education, Youth and Sports of the Czech Republic under the grant agreement LK11221, core research funding, grant GAUK 227007, and also partially supported by SVV project number 260 224. This work has been using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). Cloud computational resources were provided by the MetaCentrum under the program LM2010005 and the CERIT-SC under the program Centre CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144.

he should say to make the system understand better. We suggest that system should take initiative because the system is typically the interlocutor that misunderstands first.

2 Prototype description

Our prototype is text to text dialogue system with simple language understanding (LU) and natural language generation (NLG) component. The system updates a belief state similar to (Dušek et al., 2014) and it select actions based on handcrafted policy.

We implemented summary action which roughly correspond to core speech acts from (Traum and Hinkelman, 1992). Each summary action generates multiple candidates with different arguments taken from belief state[1]. In addition to arguments, we explicitly store human-understandable *why features*[2] which should explain to user why this action is good candidate.

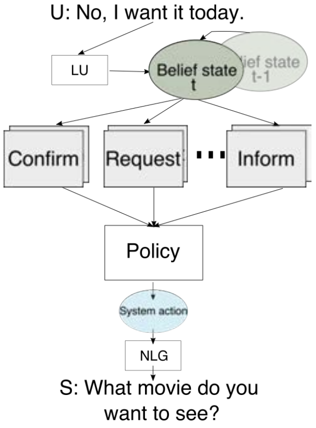


Figure 1: Architecture of our system.

The *why features* together with the system description from our knowledge base (KB) are used when the system explains the user what it tried to achieve. Our KB in addition to typical domain knowledge stores information that: a) Our system selects one from actions *Inform, Confirm, Request, Ask, Reject, Hello, and Good Bye* with corresponding arguments. b) Updates what user said in internal state.

The system also dynamically updates KB about the dialogue history so the system can use the KB as the only source of information. Having such dynamic KB allows the system not only to talk about given domain, but also about its actions, and the dialogue history without additional summary actions.

3 Discussion and Future work

To our knowledge, recovering from miscommunication received a little attention (Skantze, 2007) in dialogue systems. We argue that the system initiative to explain its action is a sensible strategy for establishing a common ground (Traum, 1994) in situation where both interlocutors are lost.

We have extended standard dialogue system architecture so the resulting system has a chance to describe its behaviour. The key addition are so called *why features*[3], the reasons for system actions. We also treat information about the system as another domain information in our KB and we do not need to implement special actions to handle help and error situations.

4 Conclusion

We presented a dialogue system architecture which allows to discuss the system behaviour easily, and thus better recovers from misunderstandings. We argued that self awareness of a system is beneficial. We plan to evaluate the system as future work.

References

Ondrej Dušek, Ondrej Plátek, Lukáš Žilka, and Filip Jurčícek. 2014. Alex: Bootstrapping a spoken dialogue system for a new domain by real users. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 79.

Gabriel Skantze. 2007. *Error handling in spoken dialogue systems: managing uncertainty, grounding and miscommunication.* Ph.D. thesis, Datavetenskap och kommunikation, Stockholm,.

David R Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599.

David R. Traum. 1994. A Computational Theory of Grounding in Natural Language Conversation. Technical report, December.

---

[1] *Confirm(movie=Tomorrow), Confirm(movie="The")*

[2] The policy does not decide based on the *why features*.

[3] The system action *confirm(date=tomorrow)* from the example conversation is stored with *why_features [user_said=tomorrow, old_value=today].*

Adaptive Dialogue Management in the KRISTINA Project for Multicultural Health Care Applications

Louisa Pragst, Stefan Ultes, Matthias Kraus, Wolfgang Minker
Ulm University
Germany
`firstname.lastname@uni-ulm.de`

Abstract

The goal of the EU-funded KRISTINA project is to help migrants in European countries get information about their resident country's health care system by the means of a socially competent dialogue system. This system has to be able to handle a considerably large dialogue domain as well as hold a natural conversation whilst taking into account the cultural background as well as the current emotional state of its dialogue partner. Dialogue management, as core component responsible for the course of the conversation, therefore needs to be able to meet these challenges. Our research is focused on adaptiveness of dialogue management to the cultural background and emotional state of the user as well as the generation of appropriate emotional responses. Furthermore the benefits of the integration of a reasoner will be investigated.

1 Introduction

In Europe, migration is ever-present. Nevertheless it can not be expected that all migrants are instantly able to speak the language of their resident country, much less to be acquainted with the culture. Under these circumstances it can be a challenge for migrants to get medical help when needed. The underlying goal of the EU-funded KRISTINA project is to provide health-related information to migrants, e.g., information about the resident country's health care system, while eliminating language and cultural barriers.

However, cultural peculiarities may hinder the interaction: elderly migrants often are reluctant in communicating health issues in a foreign environment in a manner they are not used to. While a regular person is usually not trained to deal with these cultural differences, the KRISTINA agent is intended to be a human-like, socially competent and cultural-aware system and a trustworthy source of the needed information.

There are many use cases for such a system: elderly migrants often are reluctant to see a doctor and suffer from social exclusion, their relatives have problems interacting with the local administration and temporal migrant care workers are confronted with isolation, a lack of professional training and communication problems with patients as well as supervision personnel.

To provide natural communication, the KRISTINA agent will be designed as a multimodal dialogue system having a dialogue manager (DM) at its core. Establishing the described kind of interaction results in the following challenges: as the system will be designed for many use cases, the domain of the dialogue is considerably large. Here, flexible structures are needed to approach this issue. Hence, we will integrate an external reasoning component into our dialogue manager. To ensure social competence, the user's cultural background as well as his emotional state will be taken into account. This can be further enhanced by generating appropriate emotional responses.

In the following section, we will give a short overview of the architecture of our proposed dialogue manager and continue to elaborate on how we intend to handle the inherent challenges.

2 Architecture of the dialogue manager

The general architecture of the KRISTINA dialogue manager can be seen in Figure 1. In order to render the system as a natural and socially competent dialogue partner, it takes into account cultural and emotional input in addition to the usual semantic input. Furthermore, the output will be augmented with an additional emotional response. Aside from that the dialogue manager gets additional information from a reasoning component to

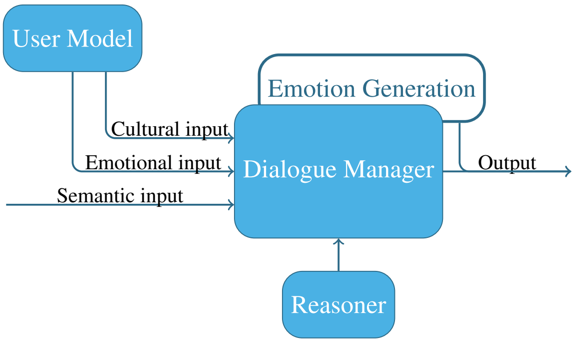


Figure 1: General architecture of our dialogue manager. In addition to semantic information it gets cultural and emotional input from the user model as well as enriched information generated by the reasoner component. The semantic output is supplemented by an appropriate emotional response.

handle the size of the dialogue domain.

The following issues have been identified as main research goals. For addressing them in our research, we will use our existing dialogue manager OwlSpeak (Ultes and Minker, 2014) and intent to explore data-driven approaches based on statistical learning methods. By that, we hope to grasp hidden aspects of human-human interaction.

2.1 Adaptation to cultural background and emotional state

For rendering the dialogue manager adaptive to multiple nationalities, the dialogue strategy will consider the cultural background of the user. This will provide migrants with a familiar way of communication and thus might help building trust. In addition, emotions may influence the dialogue strategy as well. By taking into account the emotions of the user (Bertrand et al., 2011) the acceptance of the system may be improved (e.g., (Jaksic et al., 2006; Partala and Surakka, 2004)).

2.2 Generation of appropriate emotional responses

Supplementary to the semantic representation of a system action from the dialogue manager, our research will investigate the generation of appropriate emotional responses. We expect that this will enhance the naturalness of the dialogue and thus improve user satisfaction.

2.3 Integration of a reasoner

For handling the huge domain, an external reasoning component will be integrated into the dialogue manager. All domain-related processing will be part of the reasoning leaving the dialogue manager itself with handling the interaction-related phenomena like grounding. The reasoner will identify the information which is missing to fulfill the user request. This results in a separation of dialogue management and dialogue domain forming a plug-in architecture: the reasoning component may be easily exchanged. This arrangement will improve the modularity and robustness of the overall system. A similar architecture has been described by Nothdurft et al. (2014).

3 Conclusion

In this work, we present three research goals of the KRISTINA project for rendering dialogue management multicultural and emotional with the goal of improving user acceptance and naturalness of the dialogue. Taking into account the cultural background and emotional state gives the user a sense of familiarity. By generating appropriate emotional responses, the dialogue system appears to be more human-like and thus improving the user experience. Using an external reasoner improves the modularity of the overall system and gives the dialogue manager the possibility to exploit advanced reasoning techniques.

Acknowledgments

This work is part of a project that has received funding from the *European Unions Horizon 2020 research and innovation programme* under grant agreement No 645012.

References

Gregor Bertrand, Florian Nothdurft, Wolfgang Minker, Harald Traue, and Steffen Walter. 2011. Adapting dialogue to user emotion-a wizard-of-oz study for adaptation strategies. In *Proc. of IWSDS 2011*, pages 285–294. Springer.

Nada Jaksic, Pedro Branco, Peter Stephenson, and L Miguel Encarnação. 2006. The effectiveness of social agents in reducing user frustration. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 917–922. ACM.

Florian Nothdurft, Felix Richter, and Wolfgang Minker. 2014. Probabilistic human-computer trust handling. In *Proc. of SIGDIAL 2014*, pages 51–59, Philadelphia, PA, U.S.A., June. ACL.

Timo Partala and Veikko Surakka. 2004. The effects of affective interventions in human–computer interaction. *Interacting with computers*, 16(2):295–309.

Stefan Ultes and Wolfgang Minker. 2014. Managing adaptive spoken dialogue for intelligent environments. *Journal of Ambient Intelligence and Smart Environments*, 6(5):523–539, August.

Toward a scary comparative corpus: The Werewolf Spoken Corpus

Laurent Prévot
Aix-Marseille Université, CNRS
LPL, UMR 7309, Aix-en-Provence, France
laurent.prevot@univ-amu.fr

Yao Yao
The Hong Kong Polytechnic University
Chinese and Bilingual Studies
Hung Hom, Hong Kong
ctyaoyao@polyu.edu.hk

Arnaud Gingold and **Bernard Bel**
Aix-Marseille Université, CNRS
LPL, UMR 7309, Aix-en-Provence, France
firstname.lastname@lpl-aix.fr

Kam Yiu Joe Chan
The Hong Kong Polytechnic University
Chinese and Bilingual Studies
Hung Hom, Hong Kong
joe.ky.chan@polyu.edu.hk

1 Introduction

Despite the large number of corpora—both written and spoken—that are currently available, there is still a lack of corpora that document dialogues involving more than two interlocutors. To our knowledge, there is no multilogue corpus recorded in different languages in order to perform comparative studies. This abstract describes the first steps in building such a multilogue comparable corpus for French and Mandarin languages.

Generally speaking, the development of multi-person conversational corpora is hindered by two major obstacles. First, it is relatively hard to elicit natural, multi-person conversations in a laboratory setting. While a two-person dialogue may be easily convened by an experimenter (often as an interviewer who proposes topics for discussion; e.g. the Buckeye corpus, it is much more difficult if not impossible to conduct a truly engaging discussion with a group of invited subjects who may or may not know each other very well. For this reason, existing multi-person conversation corpora mostly use recordings from naturally-occurring group meetings (e.g. research group meetings, business conferences, etc.; see ISL, ICSI (Janin et al., 2003) and AMI (Carletta et al., 2006) meeting corpora), in which case the genre of speech is limited to professional conversations that happen in work places.

The second challenge is posed by the technical difficulty of recording multi-person conversations. For one thing, group meetings are usually recorded in their natural environment, i.e. regular conference rooms, which are often not soundproof, and the speech data are collected by a few microphones placed in different spots of the room. As a result, the recordings may contain ambient noise, and it is hard to separate different talkers' speech, especially when two or more talkers speak at the same time. In addition, even if it is possible to record in a sound-treated room with multiple recording devices one set for each talker ideally, there must also be some master control mechanism that ensures that the separate tracks of recordings can be aligned properly later, but such mechanism is rarely provided in regular conference rooms.

Given the above problems, we propose to create a new spoken corpus that uses a party game to elicit multi-person conversations in a non-professional, social setting.

Moreover, one of our objective is to perform comparative quantitative analysis of this kind of data.

2 Elicitation protocol

The game we use is the Werewolf game (also known as the *Mafia* game), which is often played by a group of people (typically more than 5) at parties. Participants of the game are randomly assigned to different roles (e.g. the "murderer", the "judge", the "innocent") by drawing from a deck of cards. Apart from the judge, participants playing other roles are supposed to keep their identities only to themselves. In each round, the undercover "murderer" can "kill" one innocent player, and those who are still alive (including the "murderer") will vote who they think is the "murderer". Thus, the innocent players task is to figure out the identity of the "murderer", while the "murderer" will try to hide their identity and direct suspicion to other players. The game reaches an end either when the "innocent" players guess correctly who is the "murderer", or when the "murderer" successfully "kills" all the "innocent" players without being caught.

The Werewolf game requires very little instrument but encourages verbal communication as participants need to exchange information and opinions in order to achieve their goals. Thus, the game is often played as an ice-breaker among new friends or a pastime among familiar friends, which makes it an ideal game for eliciting natural and informal multi-person conversations among familiar or unfamiliar participants in a laboratory setting. The nature of the game involving deception and persuasion also makes it a perfect venue for studying these phenomena in conversations.

All the participants are sitting on chairs forming circle. Such a setting has the advantage of naturalness but despite usage of headset-like microphone, we still face a relatively high level of spill. The game master was recorded only for the Mandarin dataset[1].

3 Basic facts about one Werewolf game

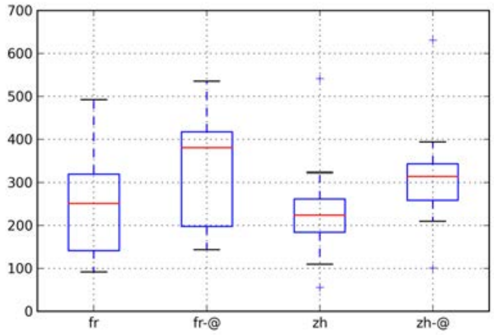


Figure 1: Actual Speaking Duration (s)

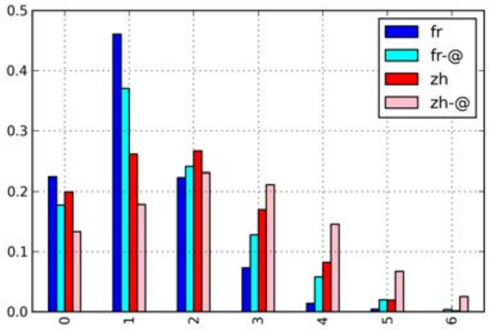


Figure 2: Number of speaker simultaneously speaking (%), sampling rate = 1Hz

For the time being we ran 4 games for each language. One game of each language has been fully transcribed. Our games lasted between 7 to 27 minutes (for an actual speech duration of 33 minutes and 3346 tokens in the later case). This important variation is partly due to games in which 'werewolves' are identified early in the game. We checked some basic properties of the sessions recorded in terms of participants involvement and simultaneously speaking. Figure 1 shows the figures for duration of speech for different speaker (excluding the game master) in the longest French and Mandarin games, including or not laughter (-@ in the figures) as verbal behavior. As expected this value is subject to strong inter-individual variations. Figure 2 provides statistics about the number of people simultaneously speaking (excluding the game master). We observe that overlapping speech amount for 35% to 65% of the data which already make it an original data set to work with.

4 Data management and curation

The WEREWOLF pilot corpus and its annotations were stored and described so as to allow a wide dissemination (permanent identifiers : `http://hdl.handle.net/11041/ortolang-000900` and `http://hdl.handle.net/11041/ortolang-000908`). Participants signed a consent form and agreed that their interactions could be publicly disseminated. The work of curation made it possible to store data in sustainable formats and facilitate reuse.

Acknowledgments

This work has been carried out thanks to the support of the A*MIDEX project (ANR-11-IDEX-0001-02) French Government program. We would like to thank all the students that participated in the recordings and in the transcription.

References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2006. The ami meeting corpus: A pre-announcement. In *Machine learning for multimodal interaction*, pages 28–39. Springer.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–364. IEEE.

---

[1] The French set-up limited the number of recorded participants to 8.

Towards a Layered Framework for Embodied Language Processing in Situated Human-Robot-Interaction

Matthias A. Priesters, Malte Schilling, Stefan Kopp
Bielefeld University, Faculty of Technology
Center of Excellence "Cognitive Interaction Technology" (CITEC)
Postfach 100 131, 33501 Bielefeld, Germany
`{mpriesters,mschilli,skopp}@techfak.uni-bielefeld.de`

Abstract

We propose a layered architecture for embodied language processing in a robot, in which the language layer is grounded in a sensorimotor layer via a schematic representation layer, which represents manual action or spatial relations in terms of embodied schemas. The schemas, on the one hand, abstract from the current sensory and motor states of the robot, and, on the other hand, enable mental simulation.

1 Introduction

The goal of the FAMULA project is to develop a bimanual robot torso able to familiarize itself with objects and their affordances.[1] This familiarization shall be scaffolded by situated dialogue grounded in unfolding manual action. Humans usually have no problem following dialogues full of underspecified references to objects or actions (e.g. containing utterances such as "no, the other one" or "yes. . . further. . . and now on top of it"), whereas for artificial agents this is no trivial task. We aim to explore the cognitive and linguistic abilities needed for the robot to engage in such situated dialogues before, during, and after action execution. The robot should be aware of its own knowledge gaps and attempt to reduce its uncertainty either by exploring the objects or by soliciting information from the tutor in situated dialogue. In such situated interaction, meaning unfolds dynamically and across language, bodily action and the interactants' environment (Goodwin, 2000), which exceeds current language-based human-robot interaction.

From the perspective of *embodied cognition*, higher-level representations are grounded in lower-level functions and are tightly interconnected (Feldman and Narayanan, 2004; Roy, 2005; Barsalou, 2008). As a consequence, cognition arises from the dynamic sensorimotor interaction with the environment. This leads to a central role of embodied representations grounded in lower-level experiences and sensorimotor behaviors. Our research focuses on modeling how embodied, situated meaning of communicative actions emerges in given interaction contexts.

2 Layered framework concept

Following the embodied cognition stance, the robot is firstly situated in a *sensorimotor* way, i.e inside its own physical form. On the one hand, its possible interactions are constrained by the shape of its hardware. On the other hand, the current states of the robot's 'body', i.e. whether it is currently engaged in an action or registers input through its sensors, influences language understanding and dialogue decision-making. Secondly, the robot is situated in its physical *environment*, i.e. in our context the scenery of objects present on the surface in front of it. These objects and their states and properties constrain the robot's options for actions and influence its needs for guidance and information. Thirdly, language use and meaning is situated in the *interaction* with the human tutor. This includes the dialogue history and common ground, the robot's interactional goals, as well as expectations and constraints for following dialogue acts or bodily actions, "interaction affordances" (Raczaszek-Leonardi et al., 2013), created by the situation and previous dialogue acts.

We devised a three-layer framework (Fig. 1) for embodied language processing. It provides higher-level representations which are embodied in the sense that they are grounded in the sensorimotor layer.

The lowest, **sensorimotor**, layer of our framework comprises the actual control primitives and

---

[1] `http://cit-ec.de/en/content/deep-familiarization-and-learning`

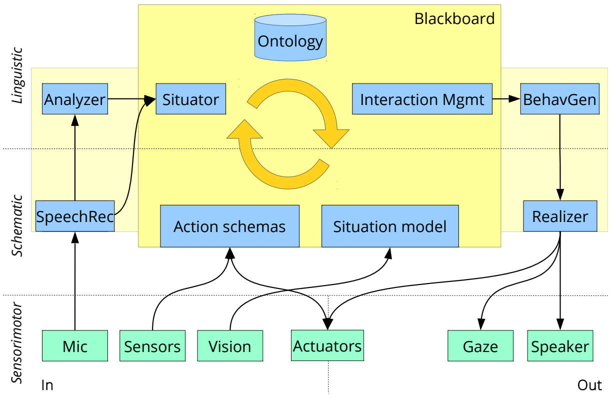


Figure 1: Proposed three-layer system architecture

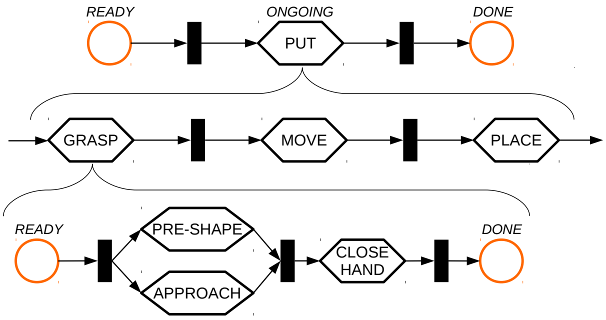


Figure 2: Hierarchical Action Schema representation for PUT (simplified)

sensor readings from the robot hardware. The hardware consists of two robotic arms mounted in a torso-like configuration onto a table. On the arms, five-digit hands are attached, which enable precise exploration and manipulation and are equipped with touch-sensitive fingertips.

To mediate between sensorimotor and linguistic layers and to realize situatedness, we introduce an intermediate **schematic** layer. Actions are represented as executable *action schemas* (Schilling and Narayanan, 2013), a Petri Net-based formalism (Fig. 2), which on the one hand represents sensorimotor states of action execution and on the other hand offers internal simulation capabilities, enabling the system to generate predictions and to represent the unfolding of actions over time. Lower-level sensory input gets accessible as part of the states of the Petri Nets. The scenery the robot is situated in is represented in a *situation model*, which keeps track of the present objects, of their ontological categories and properties.

The highest, **linguistic**, layer deals with speech input and generating answers, as well as with decision-making regarding communicative interaction. The language *analyzer* syntactically and semantically parses the input utterance using an Embodied Construction Grammar parser (Bryant, 2008). The purpose of the *situator* is situated language understanding, i.e. identifying dialogue acts and resolving references to objects or actions based on situational and ontological knowledge and on bodily and interaction states. The *interaction manager* is the decision-making component, which maintains knowledge about the level of certainty or uncertainty of the system in the current situation and decides on appropriate actions.

Implementation of the framework is work in progress and currently focuses on including infomation about the execution status of actions into language processing. For example, when resolving an ambiguous, underspecified reference ("no, the other one"), the system takes into account which object is currently in focus (e.g. the moved object vs. the target location of a PUT action) depending on the state of the ongoing action (GRASP vs. MOVE/PLACE, Fig. 2).

Acknowledgments

This work is supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

References

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

John Bryant. 2008. *Best-fit constructional analysis*. Ph.D. thesis, University of California, Berkeley.

Jerome Feldman and Srinivas Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392.

Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522.

Joanna Raczaszek-Leonardi, Iris Nomikou, and Katharina J. Rohlfing. 2013. Young children's dialogical actions: The beginnings of purposeful intersubjectivity. *IEEE Trans Auton Ment Dev*, 5(3):210–221.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1–2):170–205.

Malte Schilling and Srini Narayanan. 2013. Communicating with executable action representations. In *AAAI 2013 Spring Symposium Series*, Stanford.

How Questions and Answers Cohere

Mandy Simons
Carnegie Mellon University, Department of Philosophy
5000 Forbes Ave
Pittsburgh, PA, 15213, USA
`simons@andrew.cmu.edu`

1 The basic observation

When a declarative sentence is uttered in response to a question, the asserted content may be richer than the compositionally derivable content for that sentence. Here are some illustrative examples:

(1) Q:What did Clara draw with her new pencil?
A: She drew a dragon.
Asserted content: Clara drew a dragon **with her new pencil**.

(2) Q: What's Jane wearing for the wedding?
A: She's wearing jeans and a t-shirt.
Asserted content: Jane is wearing jeans and a t-shirt **for the wedding**.

(3) Q: What's Harriet knitting for Henry?
A: She's knitting a scarf.
Asserted content: Harriet is knitting a scarf **for Henry**.

I provide here an account of the phenomenon in DRT (Kamp 1981), enriched with ideas from SDRT (Asher 1993). I posit a discourse relation *Direct Answer* (DirAns), which I take to be presumed whenever a question is followed by an assertoric response, and propose that the introduction of this relation triggers a special update rule which results in merging the contents of the question and its answer.

2 Representing questions

I assume that at some level of representation, a *wh*-question has the form shown in (4):

(4) $[wh_1 \ldots [\, wh_n \, [_S \ldots x_1 \ldots x_n \ldots \,]]$

The embedded S can be treated by standard DRS construction rules, except for the *wh*-traces, for which I propose the rule in (5). (The DRS which results will be embedded under a λ-operator; I omit those details here. Cf. the treatment of questions in Krifka 2001.)

(5) **DRS construction rule for *wh*-traces**
Given the syntactic configuration: $[_{XP}\ x_i{:}\ \varphi_1 \ldots \varphi_n\ ]$, ($\varphi_1 \ldots \varphi_n$ being the semantic type features derived from the *wh*-expression itself), where x_i is bound by a *wh*-expression:

i. introduce a new discourse referent x_i into the universe of the DRS under construction, and conditions $\varphi_1(x_i) \ldots \varphi_n(x_i)$ into the set of conditions of that DRS.

ii. Then, add a condition of the form , $?x_i$ to the set of conditions of the DRS.

The condition $?x_i$ marks x_i as a *forward looking anaphor*, indicating that some new predication containing information about this d-ref is anticipated in the subsequent discourse. I propose that in order to maintain the relation DirAns between Q and S, update with S *must* provide this information: call this the *Answerhood Constraint*.

3 Providing a value does not suffice

Observing that an answer must provide a value for the forward-looking *wh*-anaphor leads to the idea that the only purpose of a full sentence answer is to provide that value. On this picture, the update rule for Direct Answers should simply "extract" the value for the *wh*-anaphor from the content provided by the answer. Two observations show that this is not correct. First: what is asserted by utterance of a declarative in response to a question may include content contained in the answer but not in the question, as shown in (6):

(6) Q:What did Clara draw with her new pencil?
A: In the morning, (she drew) a dragon, and in the afternoon, (she drew) a snake.

This shows that the full content of the answer is semantically relevant: we arrive at the interpretation by *combining* the content of the question and the content of the answer.

Second: to count as an answer, it is not sufficient for a response to provide a potential value for the *wh*-anaphor. Consider (7):

(7) Q: Who did Jane see?
A: Frankie loves [$_F$ Billie].

The focus marking in (7)A unambiguously marks *Billie* as intended to provide the value for the *wh*-anaphor; but lack of match between contents of question and assertion render this unsuitable as an answer. Clearly, the content of the answer matters: for an utterance to count as a direct answer it must be construable as being, roughly, "about" the same thing as the question.

The first observation suggests that we should construct the content of answers by *merging* the linguistically expressed contents of question and answer. The second observation suggests that in this process of merge, we should seek to *unify* the content of the question with the content of the answer where possible.[1] The Answerhood Constraint requires that this procedure should provide an answer to the question.

4 Merge + Unification

The Merge+Unification procedure is triggered by introduction into the SDRS (omitted here) of the condition DirAns(Q,A), where Q, A are discourse segments.

Merge
If DirAns(Q,A), then revise K(A) to K(Q+A) as follows:

i. $U(K(Q+A)) = U(K(Q)) \cup U(K(A))$
ii. $Con(K(Q+A)) = Con(K(Q)) \cup Con(K(A))$

Unify
$\forall x \in U(K(Q+A))$, if $\exists y \in U(K(Q+A))$ s.t. positing $x=y$ does not lead to inconsistency, then add $x=y$ to Con(K(Q+A)).

Example

(8) Q: What did Clara draw with her new pencil?
A: In the morning, she drew a dragon.

i. K_Q: λx_3 [e_1, x_1, x_2, x_3 : x_1=Clara, her-new-pencil(x_2), draw(e_1), Ag(e_1, x_1), Instr(e_1, x_2), Th(e_1, x_3), non-person(x_3), ?x_3]
ii. K_A: [e_2, y_1, y_2: female(y_1), y_1=?, draw(e_2), Ag(e_2, y_1), Th(e_2, y_2), dragon(y_2), $e_2 \subseteq$the-morning]
iii. Assume: DirAns(Q,A)
iv. Merge: Revise K_A to K(Q+A): [e_1, x_1, x_2, x_3, e_2, y_1, y_2: x_1=Clara, her-new-pencil(x_2), draw(e_1), Ag(e_1, x_1), Instr(e_1, x_2), Th(e_1, x_3), non-person(x_3), ?x_3, female(y_1), y_1=?, draw(e_2), Ag(e_2, y_1), Th(e_2, y_2), dragon(y_2), $e_2 \subseteq$the-morning]
v. Unify: [e_1, x_1, x_2, x_3, e_2, y_1, y_2: x_1=Clara, her-new-pencil(x_2), draw(e_1), Ag(e_1, x_1), Instr(e_1, x_2), Th(e_1, x_3), non-person(x_3), female(y_1), draw(e_2), Ag(e_2, y_1), Th(e_2, y_2), $e_2 \subseteq$the-morning, dragon(y_2), e_1=e_2, x_1=y_1, x_3=y_2]

Here, it is consistent to identify e_1 and e_2, as both are drawing events, and the information about participants is compatible. This forces us to identify the *wh*-anaphor with y_2 (the dragon), satisfying the Answerhood Constraint. More complex examples will require us to further elaborate and refine the unification procedure.

The account does not yet solve the problem posed by (7), where the *wh*-anaphor could be identified with the d-ref corresponding to *Billie* without inconsistency. However, whereas in (7), identification of these d-refs is merely allowable, in the felicitous (8), identification is necessitated by the overall pattern of unification of referents. I propose that this is what is required to satisfy the Answerhood Constraint; mere consistency does not suffice.

As a side-benefit, this approach allows for a straightforward characterization of direct answers: utterances whose interpretation results in satisfaction of the Answerhood Constraint. No restrictions on the form of the utterance or the logical form of its content are required.

5 References

Asher, N. 1993. *Reference To Abstract Objects in Discourse.* Kluwer Academic Publishers, Dordrecht.

Hobbs, J. R., Stickel, M. E., Appelt, D. E, and Martin, P. 1993. Interpretation as abduction. *Artificial Intelligence* 63: 69-142.

Kamp, Hans 1981. A theory of truth and semantic representation. In J. Groenendijk, TH. Janssen and M. Stokhof (eds.) *Formal Methods in the Study of Language*, Part 1. Mathematisch Centrum. Amsterdam: 277-322.

Krifka, M. 2001. For a structured meaning account of questions and answers. C. Féry and W. Sternefeld (eds), *Audiatur vox sapientiae: A festschrift for Arnim von Stechow*.Akademie Verlag, Berlin: 287-319.

---

[1] This is in accord with the principle of Hobbs et al. 1993 to eliminate redundancies wherever possible.

Modeling Referential Coordination as a Particle Swarm Optimization Task

H. Chase Stevens
University of Edinburgh
3 Charles Street
Edinburgh, UK
`chase@chasestevens.com`

Hannah Rohde
University of Edinburgh
3 Charles Street
Edinburgh, UK
`hannah.rohde@ed.ac.uk`

Abstract

We take a novel approach to modeling the influence of production cost on referential coordination by employing particle swarm optimization (PSO), a general-purpose optimization method. The PSO-based model replicates behaviors observed in previous research (Rohde et al., 2012; Brennan & Clark, 1996), indicating that referential coordination can be framed as a constrained optimization problem in which agents may need only to maintain a simplified representation of the common ground.

1 Introduction

The question of how referential choice and interpretation are influenced by production cost remains unresolved in the literature. When producing referring expressions under potentially ambiguous conditions, speakers must weigh the cost of producing an expression against the ease with which their conversational partners will be able to infer the intended referent. Recent research (Rohde et al., 2012; Degen & Franke, 2012; Frank & Goodman, 2012) investigates the conditions under which speakers coordinate the use of ambiguous expressions through the use of language games; in Rohde et al., an iterated language game was introduced which targeted dyadic referential coordination over multiple turns. A study conducted using this game showed that the participants' ability to successfully coordinate the use of less costly ambiguous forms was sensitive to the relative cost of competing unambiguous forms.

We build a computational model in order to simulate the findings of said study and, in doing so, to better understand the relationship between form costs and referential coordination. PSO, a swarm-based, general-purpose global optimization method, was chosen for this simulation as it allowed for a natural representation of agents' interactions over time as they sought to jointly optimize their performance in the language game. Our PSO model is capable both of replicating the coordination behaviors observed in human participants and of extrapolating beyond the conditions investigated in the Rohde et al. study.

2 Methods

PSO represents potential solutions for maximizing an objective function as particle positions existing within a multidimensional space. Over a number of iterations, increasingly suitable solutions are found as particles explore this space, their paths influenced by the thus far best-found solutions (Kennedy & Eberhart, 1995).

To adapt PSO to the Rohde et al. language game, we model agents' strategies as sets of probabilities, yielding a search space in which each dimension represents the preference of an agent to use an ambiguous referring expression for some referent instead of an unambiguous expression of differing production cost. In this way, a pair of particles can represent two interlocutors with two individual referential strategies, the fitness of which can be evaluated with regards to the production costs and successful communication rewards imposed by the language game in conjunction with the likelihood of successful communication as dictated by the partner's referential strategy.

Whether two agents ultimately succeed in coordinating (and on which referring expressions) is, under this approach, a function of whether the agents' strategies after the final PSO iteration are consistent and compatible with each other. To best replicate the influences of form cost on coordination as observed in Rohde et al., we optimize the parameters of the PSO algorithm, which specify the movement of particles, to the study.

3 Results and discussion

Our PSO-based model, parameterized using optimized values, is able to successfully capture the relative effects of varying form costs on referential coordination rate observed by Rohde et al.. Additionally, when agents within the PSO simulation are forced to entrain on the use of a high-cost unambiguous form, then subsequently, due to a change in discourse context, are permitted the use of a less costly, previously ambiguous form, they continue to make use of the form on which they have entrained; in doing so, agents in our model capture the overinformativity behavior noted in Brennan and Clark.

In extrapolating beyond the conditions of the Rohde et al. study, our PSO model predicts the likelihood of coordination on the ambiguous form to increase in response to both higher successful communication reward and higher ambiguous form cost. The former prediction is in keeping with the conclusions of Rohde et al.; the latter follows given that higher ambiguous form costs, in exceeding the costs of competing unambiguous forms, reduce the number of referents for which using the ambiguous form is beneficial and, in this way, reduce the number of competing viable referential strategies.

Agents within the PSO model do not maintain explicit representation of the common ground or of their communicative partners. Instead, agents consist only of a position within the search space of referential strategies and a velocity through that space. We interpret the model's success in replicating human referential coordination to be in keeping with more egocentric models of communication (Horton & Keysar, 1996) than audience design views, especially given that agents maintain no explicit model of their communicative partners or of the common ground.

4 Conclusion

Our PSO-based modeling technique captures and extrapolates beyond experimentally observed behavior, enabling exploration of the influence of form costs on referential coordination and referring expression production.

Further, our application of PSO to modeling referential coordination demonstrates not only that referential coordination in humans can be explained in terms of a generalized optimization process, but also suggests a lower bound for how complex agents need to be in order to respond to form costs in a manner similar to humans.

References

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Degen, J., & Franke, M. (2012). Optimal reasoning about referential expressions. *Proceedings of SemDIAL.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998-998.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91-117.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *IEEE international conference on neural networks, 1995. Proceedings.* (Vol. 4, p. 1942-1948). doi: 10.1109/ICNN.1995.488968

Rohde, H., et al. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In *The 16th workshop on the semantics and pragmatics of dialogue.* Paris.

Distribution of Non-Sentential Utterances across BNC Genres: A Preliminary Report

Kwong-Cheong Wong
Université Paris-Diderot (Paris 7)
Paris, France
wongkwongcheong@gmail.com

Jonathan Ginzburg
Université Paris-Diderot (Paris 7)
Paris, France
yonatan.ginzburg@univ-paris-diderot.fr

1 Introduction

One of the most important features of conversation is that it is genre-based. It is the distinctive framework of a specific genre that establishes the normative expectations which underpin, and which actually drive, the interaction in a conversation. Ginsburg (2012), building on the works of others, formulates a theory of meaning for non-sentential utterances (NSUs) and gives a rudimentary treatment of genre specificity. A step forward related to incorporating fully-fledged genre specificity into that theory of meaning is to study the distribution of NSUs across different genres and to identify the distinctive characteristics of NSUs in each distinct genre. This paper is a preliminary report on an on-going empirical study concerning NSUs in the genres of the British National Corpus (BNC).

2 Method and Findings

There are totally 23 (classified) genres in the 10-million-word spoken part of BNC (see the left column of Table 1); for the rationale behind this categorization, see Lee (2002). The sizes of these genres vary greatly, ranging from 15,105 words of *S_lect_commerce* (lecture on commerce) to 4,206,058 words of *S_conv* (causal conversation). In order to cover the NSU phenomena as many as possible in those smaller genres and to make the study feasible, files of total size in the range of 15,000-19,999 words are randomly selected from each genre. The selected sub-corpus consists of 69 files, totalling 383,979 words.

Table 1 below shows the frequency distribution of NSUs across different BNC genres in terms of percentage which is calculated, for each genre, by the formula: (total number of NSUs / total number of sentence units) x 100%. As to be expected, those genres which are more interactive in nature (e.g., interview, medical consultation, classroom, causal conversation, demonstration, and broadcast discussion) have high frequencies of NSUs (more than 20%), whereas those genres which are not interactive in nature (e.g., broadcast news, parliament, and sermon) have few NSUs.

BNC Genre	Frequency of NSUs
S_interview_oral_history (Hist)	33.6%
S_consult (Cons)	27.8%
S_classroom (Class)	27.2%
S_interview (Intv)	26.6%
S_conv (Conv)	25.7%
S_demonstratn (Demo)	22.7%
S_brdcast_discussn (Discn)	22.4%
S_meeting (Meet)	15.8%
S_courtroom (Court)	15.4%
S_lect_humanities_arts (H_arts)	14.5%
S_tutorial (Tut)	14.3%
S_pub_debate (P_deb)	13.3%
S_speech_unscripted (Sp_us)	13.1%
S_lect_polit_law_edu (P_law)	8.2%
S_speech_scripted (Sp_s)	6.1%
S_lect_nat_science (Nat_sc)	6.0%
S_lect_soc_science (Soc_sc)	5.1%
S_brdcast_documentary (Doc)	2.0%
S_sportslive (Sport)	1.7%
S_lect_commerce (Comm)	1.2%
S_brdcast_news (News)	0%
S_parliament (Prlmnt)	0%
S_sermon (Sermn)	0%

Table 1: Distribution of NSUs across BNC Genres

It is interesting, however, to observe that there is a significant variation of NSUs among the five lecture genres, ranging from 1.2% of *S_lect_ commerce* to 14.5% of *S_lect_humanities_arts*.

This study is based on the NSU taxonomy given by Fernández and Ginsburg (2002), which consists of 15 classes of NSUs covering various kinds of acknowledgments (plain acknowledgement, repeated acknowledgement), questions (clarification ellipsis, sluice, check question), answers (short answer, plain affirmative answer, repeated affirmative answer, propositional modifier, plain rejection, helpful rejection), and extensions (factual modifier, bare modifier phrase, conjunction + fragment, filler). Tables 2, 3 and 4 below show the composition of NSUs in each BNC genre.

	Ack	RepAck	CE	Sluice	CheckQ
Hist	25.50%	1.80%	0.70%	0.00%	0.20%
Cons	18.90%	0.70%	1.10%	0.10%	0.60%
Class	14.90%	2.60%	0.50%	0.30%	1.00%
Intv	20.50%	0.80%	0.60%	0.00%	1.30%
Conv	10.70%	0.80%	3.60%	0.70%	0.40%
Demo	9.40%	1.10%	0.10%	0.20%	0.60%
Discn	15.30%	0.90%	0.70%	0.40%	0.20%
Meet	10.20%	0.70%	0.40%	0.10%	0.50%
Court	10.10%	0.60%	0.40%	0.00%	0.00%
H_arts	10.60%	0.50%	0.40%	0.10%	0.00%
Tut	9.20%	0.40%	0.70%	0.00%	0.00%
P_deb	8.90%	0.10%	0.30%	0.00%	0.00%
Sp_us	5.20%	1.60%	0.70%	0.00%	0.20%
P_law	0.80%	1.60%	0.60%	0.00%	0.00%
Sp_s	3.10%	0.80%	0.40%	0.00%	0.00%
Nat_sc	1.60%	0.10%	0.30%	0.00%	1.60%
Soc_sc	2.30%	0.80%	0.10%	0.10%	0.00%
Doc	0.60%	0.10%	0.40%	0.20%	0.00%
Sport	0.80%	0.30%	0.00%	0.00%	0.00%
Comm	0.70%	0.20%	0.00%	0.00%	0.00%
News	0.00%	0.00%	0.00%	0.00%	0.00%
Prlmnt	0.00%	0.00%	0.00%	0.00%	0.00%
Sermn	0.00%	0.00%	0.00%	0.00%	0.00%

Table 2: Composition of NSUs in BNC Genres

	ShortAns	AffAns	RepAffAns	PropMod	Reject
Hist	0.80%	1.80%	0.40%	0.10%	0.70%
Cons	0.60%	3.00%	0.50%	0.10%	1.30%
Class	4.00%	1.70%	0.20%	0.10%	0.70%
Intv	0.10%	1.40%	0.10%	0.40%	0.30%
Conv	1.70%	4.50%	0.20%	0.20%	1.70%
Demo	5.50%	2.60%	0.40%	0.10%	1.50%
Discn	1.60%	1.40%	0.20%	0.40%	0.90%
Meet	0.60%	1.80%	0.10%	0.30%	0.40%
Court	1.10%	1.30%	0.20%	0.40%	1.00%
H_arts	0.40%	1.90%	0.00%	0.20%	0.20%
Tut	0.60%	1.70%	0.50%	0.10%	0.40%
P_deb	0.30%	2.30%	0.20%	0.30%	0.60%
Sp_us	2.80%	0.90%	0.20%	0.20%	0.50%
P_law	3.90%	0.20%	0.20%	0.00%	0.00%
Sp_s	0.20%	0.90%	0.00%	0.00%	0.30%
Nat_sc	1.40%	0.90%	0.00%	0.00%	0.00%
Soc_sc	0.40%	0.60%	0.00%	0.00%	0.30%
Doc	0.20%	0.10%	0.10%	0.00%	0.20%
Sport	0.10%	0.00%	0.10%	0.00%	0.10%
Comm	0.00%	0.20%	0.00%	0.00%	0.00%
News	0.00%	0.00%	0.00%	0.00%	0.00%
Prlmnt	0.00%	0.00%	0.00%	0.00%	0.00%
Sermn	0.00%	0.00%	0.00%	0.00%	0.00%

Table 3: Composition of NSUs in BNC Genres (cont.)

	HelpReject	FactMod	BareModPh	Conj+Frag	Filler
Hist	0.20%	0.40%	0.20%	0.40%	0.50%
Cons	0.30%	0.30%	0.10%	0.00%	0.40%
Class	0.20%	0.40%	0.00%	0.00%	0.50%
Intv	0.00%	0.60%	0.00%	0.40%	0.30%
Conv	0.20%	0.80%	0.10%	0.00%	0.10%
Demo	0.00%	0.20%	0.00%	0.10%	0.90%
Discn	0.00%	0.00%	0.20%	0.10%	0.10%
Meet	0.00%	0.10%	0.00%	0.10%	0.40%
Court	0.10%	0.20%	0.00%	0.00%	0.00%
H_arts	0.20%	0.00%	0.00%	0.00%	0.00%
Tut	0.10%	0.40%	0.00%	0.00%	0.20%
P_deb	0.10%	0.20%	0.00%	0.00%	0.20%
Sp_us	0.00%	0.40%	0.20%	0.00%	0.10%
P_law	0.20%	0.80%	0.00%	0.00%	0.00%
Sp_s	0.10%	0.00%	0.00%	0.10%	0.00%
Nat_sc	0.00%	0.00%	0.00%	0.00%	0.00%
Soc_sc	0.10%	0.10%	0.00%	0.00%	0.10%
Doc	0.00%	0.10%	0.00%	0.00%	0.00%
Sport	0.10%	0.00%	0.10%	0.10%	0.10%
Comm	0.00%	0.00%	0.00%	0.00%	0.00%
News	0.00%	0.00%	0.00%	0.00%	0.00%
Prlmnt	0.00%	0.00%	0.00%	0.00%	0.00%
Sermn	0.00%	0.00%	0.00%	0.00%	0.00%

Table 4: Composition of NSUs in BNC Genres (cont.)

It can be observed that, except for *S_lect_polit_law_edu* (and those genres which have no NSUs), *Ack* (plain acknowledgement) is the most frequent among all NSUs in each genre. Apart from this general observation, some distinctive characteristics of some genres can also be observed. For example, *S_classroom* has a high frequency of *ShortAns* due to the fact that students frequently give answers to the teacher's questions; *S_cons* (medical consultation) has a high frequency of *AffAns* and *Reject* due to the fact that patients frequently give yes-or-no answers to the doctor's probing; *S_conv* (causal conversation), due to its causal nature, has a high frequency of *CE* (clarification ellipsis).

3 Further Work

A more in-depth study on a few selected BNC genres will be conducted.

References

Fernández, R., & Ginzburg, J. (2002). Non-sentential utterances: A corpus-based study. *Traitement automatique des languages*, 43(2), 13-42.

Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*, Oxford University Press.

Lee, D. (2002). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language and Computers*, 42(1), 247-292.

Interactive Learning through Dialogue for Multimodal Language Grounding

Yanchao Yu
Interaction Lab
Heriot-Watt University
Edinburgh,UK
`yy147@hw.ac.uk`

Oliver Lemon
Interaction Lab
Heriot-Watt University
Edinburgh,UK
`o.lemon@hw.ac.uk`

Arash Eshghi
Interaction Lab
Heriot-Watt University
Edinburgh,UK
`arash.eshghi@hw.ac.uk`

Abstract

We present initial work addressing the problem of interactively learning perceptually grounded word meanings in a multimodal dialogue system. We design an incremental dialogue system using Type Theory with Records (TTR) semantic representations for learning about visual attributes of objects through natural language interaction. This paper explores the use of multi-label visual attribute classification models (TRAM and MLKNN) for such a system. However, these models are found not to perform adequately for this task, so we suggest future directions.

1 Introduction

Learning to identify and talk about objects/events in the surrounding environment is a key capability for intelligent, goal-driven systems what interact with other agents and external world, e.g. smart phones and robots. There has recently been a surge of works and significant progress made on generating image descriptions, identifying images/objects using text descriptions, as well as classifying/describing novel objects using low-level concepts (e.g. colour and shape) (Farhadi et al., 2009). However, most systems rely on pre-trained data of high quality and high quantity without possibility of online error correction. Furthermore, they are unsuitable for robots and multimodal systems that continuously, and incrementally learn from the environment, and may encounter objects they haven't seen in training data. These limitations may be alleviated if systems can learn concepts from situated dialogue with humans. NL interaction enables systems to take initiative and seek the particular information they need or lack by e.g. asking questions with the highest information gain (see e.g. (Skocaj et al., 2011), and Fig. 1).

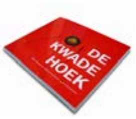
Dialogue	**Image**	**Final semantics**
S: Is this a green mug? T: No it's red S: Thanks.		$\begin{bmatrix} x_{=o1} & : & e \\ p2 & : & red(x) \\ p3 & : & mug(x) \end{bmatrix}$
T: What can you see? S: something red. What is it? T: A book. S: Thanks.		$\begin{bmatrix} x1_{=o2} & : & e \\ p & : & book(x1) \\ p1 & : & red(x1) \\ p2 & : & see(sys, x1) \end{bmatrix}$

Figure 1: Example dialogues & resulting semantic representations

We present the first step in a larger programme of research with aim of developing dialogue system what learns (visual) concepts – word meaning – through situated dialogues with humans. We integrate a basic dialogue system using DS-TTR (Eshghi et al., 2012), with two multi-label classification models (MLkNN and TRAM) to simulate the interactive learning process. In effect, the dialogue with a tutor continuously provides semantic information about objects in the scene which is then fed to an online classifier in the form of training instances. Conversely, the system can utilise the grammar and existing knowledge base to make references and formulate questions related to different objects' attributes identified in the scene. For evaluating the performance of situated dialogue on attribute-based recognition, we compare the performance of two learning models as more training instances are presented to them.

2 System Architecture

The architecture of the system (see Fig. 2) contains two main modules: a vision module for visual feature extraction and classification; and a dialogue system module using DS-TTR. We assume access to logical semantic representations by DS-

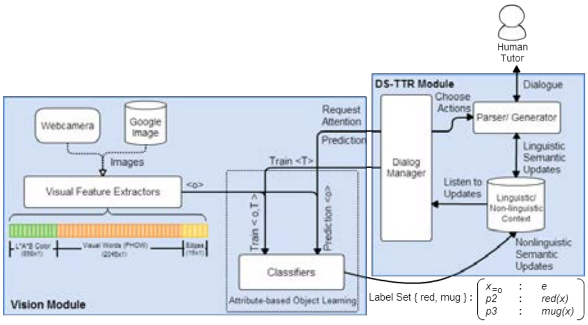


Figure 2: Architecture of the teachable system

TTR parser/generator as a result of processing dialogues with a human tutor.

The **Vision Module** is implemented with two multi-label classification algorithms (MLkNN and TRAM) for learning/classifying low-level object-based attributes: a) MLkNN (Zhang and Zhou, 2007), as a supervised learning model, predicts potential label sets of unseen objects using k-nearest neighbour algorithm; b) TRAM (Kong et al., 2013) proposes a semi-supervised model that predicts the binary label set of a novel instance based on utilized information from both seen and unseen objects. For learning new multi-label classifiers, we build a pair of inputs – a 2960-dimensional visual feature vector from each object using features from (Farhadi et al., 2009) and an i-dimensional binary label vector for each instance (where the $i - th$ attribute takes the value of 1 if it belongs to the instance and -1 otherwise).

The **Dialogue System Module** implements DS-TTR, which is a word-by-word incremental semantic parser/generator for dialogue, based around the Dynamic Syntax (DS) grammar framework (Cann et al., 2005), in which interlocutors interactively construct contextual and semantic representations (Purver et al., 2011). The contextual representations afforded by DS-TTR are of the fine-grained semantic content that is jointly negotiated/agreed upon by the interlocutors, as a result of processing questions and answers, clarification requests, corrections, acceptances, etc (see (Eshghi et al., 2015) and the first row of Fig. 1).

3 Results & Future work

We evaluated the performance of two multi-label classification models (MLkNN and TRAM) for attribute classification of object images. TRAM outperforms MLkNN and both models improve on classifying attributes for which they receive more training instances. However, the results show that both models are not ideal approaches to our problem, since for good performance they require many more training examples than can be provided in an interactive teaching session with a human. What we need are learning methods which can operate effectively on small numbers of samples, and which can improve performance robustly while continuously learning new examples. These properties are know as “zero-shot” and “incremental” learning respectively. We will explore these two approaches in future work.

Acknowledgments

This research is partially supported by the EPSRC, under grant number EP/M01553X/1 (BABBLE project[1]).

References

Ronnie Cann, Ruth Kempson, and Lutz Marten. 2005. *The Dynamics of Language*. Elsevier, Oxford.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguisitics.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*.

Xiangnan Kong, Michael K. Ng, and Zhi-Hua Zhou. 2013. Transductive multilabel learning via label set propagation. *IEEE Trans. Knowl. Data Eng.*, 25(3):704–719.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In J. Bos and S. Pulman, editors, *Proceedings of the 9th International Conference on Computational Semantics*, pages 365–369, Oxford, UK, January.

Danijel Skocaj, Matej Kristan, Alen Vrecko, Marko Mahnic, Miroslav Janícek, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 3387–3394.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048.

---

[1] `https://sites.google.com/site/hwinteractionlab/babble`