SEMDIAL 2013
DialDam

Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue

Raquel Fernández and Amy Isard (eds.)

Amsterdam, 16–18 December 2013

UNIVERSITY OF AMSTERDAM

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

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

DialDam Website

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

DialDam Sponsors

Netherlands Organisation
for Scientific Research

Benelux Association
for Artificial Intelligence

INSTITUTE FOR LOGIC,
LANGUAGE AND COMPUTATION

Gemeente
Amsterdam

DialDam Endorsements

SIGdial

SIGSEM

Preface

DialDam brings the SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue back to Amsterdam, where the second meeting — Amstelogue — took place in 1999. The return to Amsterdam has brought about the opportunity to collocate SEMDIAL with the Amsterdam Colloquium, a premier forum for formal work on semantics and pragmatics. We are excited about this opportunity and we hope that it will encourage interaction between the two communities, which seem to have drifted somewhat apart over the last few years. To stimulate contact, the last session of DialDam will be offered as a special session on dialogue within the Amsterdam Colloquium.

While in the initial years of the SEMDIAL series the focus was on what could be called classic semantics and pragmatics, over the last decade and a half the scope of SEMDIAL has broadened significantly. It now covers a wide range of topics concerned with aspects of dialogue, ranging from the acquisition of conversational competence by children, experimental semantics and psycholinguistic studies, to work on the design of artificial conversational agents, human interaction with robots, and the computational modelling of disfluencies, gesture, gaze, and turn taking. The collection of papers presented in this volume offer a good overview of the diversity of angles from which the study of dialogue can be tackled.

We received a total of 40 full paper submissions, 17 of which were accepted after a peer-reviewing process during which each submission was reviewed by three experts. We are extremely grateful to the Programme Committee members for their very detailed and helpful reviews. Three papers amongst the accepted submissions were selected for presentation in the special session with the Amsterdam Colloquium, in collaboration with Maria Aloni (representing the Amsterdam Colloquium) and on the basis of affinity with the scope of the colloquium. In response to a later call, we received a total of 30 abstract submissions describing ongoing projects or system demonstrations, of which 26 were accepted for poster presentation. Abstract submissions were not refereed, but evaluated for relevance to the SEMDIAL topics of interest only by ourselves in our role as chairs. All accepted full papers and poster abstracts are included in this volume.

In addition, the DialDam programme features three keynote presentations by Danielle Matthews, Marc Swerts, and Matthew Stone (who will give a plenary talk for the audiences of both SEMDIAL and the Amsterdam Colloquium). We are honoured to have them at the workshop and are looking forward to their talks. Abstracts of their contributions are also included in this volume.

DialDam has received generous financial support from the Netherlands Organisation for Scientific Research (NWO), the Benelux Association for Artificial Intelligence (BNVKI), the Gemeente Amsterdam, and the Institute for Logic, Language and Computation (ILLC), which hosts the event. We are very grateful for their sponsorship, as well as for the endorsement by the ACL Special Interest Groups SIGdial and SIGSEM; and last but not least for the tireless work of Inés Crespo who helped with all aspects of the local organisation, as well as of Peter van Ormondt from the ILLC office.

Raquel Fernández and Amy Isard

Amsterdam & Edinburgh

December 2013

Programme Committee

Ron Artstein	Institute for Creative Technologies USC
Ellen Gurman Bard	University of Edinburgh
Luciana Benotti	Universidad Nacional de Córdoba
Claire Beyssade	CNRS/Institut Jean Nicod, Paris
Nate Blaylock	Nuance Communications
Holly Branigan	University of Edinburgh
Susan Brennan	Stony Brook University
Sarah Brown-Schmidt	University of Illinois
Mark Core	University of Southern California
Paul Dekker	ILLC, University of Amsterdam
David Devault	USC Institute for Creative Technologies
Simon Dobnik	Gothenburg University
Jens Edlund	Royal Technical Institute (KTH)
Raquel Fernández (**chair**)	ILLC, University of Amsterdam
Victor Ferreira	UC San Diego
Claire Gardent	CNRS/LORIA Nancy
Kallirroi Georgila	University of Southern California
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Amy Isard (**chair**)	University of Edinburgh
Andrew Kehler	UC San Diego
Ruth Kempson	Kings College London
Emiel Krahmer	Tilburg University
Staffan Larsson	Gothenburg University
Alex Lascarides	University of Edinburgh
Oliver Lemon	Hariot-Watt Univeristy
Colin Matheson	University of Edinburgh
Danielle Matthews	University of Sheffield
Gregory Mills	University of Edinburgh
Chris Potts	Stanford University
Laurent Prévot	Laboratoire Parole et Langage
Matthew Purver	Queen Mary University of London
Hannes Rieser	Bielefeld University
David Schlangen	Bielefeld University
Gabriel Skantze	Royal Technical Institute (KTH)
Amanda Stent	Stony Brook University
Matthew Stone	Rutgers
David Traum	Institute for Creative Technologies USC
Nigel Ward	University of Texas at El Paso
Henk Zeevat	ILLC, University of Amsterdam

Table of Contents

Invited Talks

The Development of Reference in Early Childhood: the Roles of Communicative Motivation, Cognitive Abilities and the Caregiving Environment 2
Danielle Matthews

Coherence and Meaning in Situated Dialogue 3
Matthew Stone

On Variability in Pitch Accent Distributions 4
Marc Swerts

Full Papers

Demand Characteristics as a Tool for Evaluating the Design of Collaborative Tasks 6
Ed Baggs

Update conditions and intensionality in a type-theoretic approach to dialogue semantics 15
Robin Copper

Visual and linguistic predictors for the definiteness of referring expressions 25
Manjuan Duan, Micha Elsner and Marie-Catherine de Marneffe

People hesitate more, talk less to virtual interviewers than to human interviewers 35
Fauren Faust and Ron Artstein

On the use of social signal for reward shaping in reinforcement learning for dialogue management 44
Emmanuel Ferreira and Fabrice Lefèvre

When Apples Mean Oranges: Lexical De-Entrainment and Alignment Theories 54
Christina Freihorst, Simón Algorta, Vivien Mast and Robert Porzel

The earliest utterances in dialogue: towards a formal theory of parent/child talk in interaction . . . 63
Jonathan Ginzburg and Sara Moradlou

English reverse prosody in responses to yes-no questions 73
Daniel Goodhue, James Pickett and Michael Wagner

Effects of Belief and Memory on Strategic Negotiation 82
Markus Guhe, Alex Lascarides, Kevin O'Connor and Verena Rieser

Modelling Expectation in the Self-Repair Processing of Annotat-, um, Listeners 92
Julian Hough and Matthew Purver

Be Open To Computer Based Coaching 102
Daniel Mäurer, Anna Bruns and Karsten Weihe

If you repeat your interlocutor's syntactic structure, you are likely to repeat her pronunciation, too 112
Marie Postma-Nilsenova, Lisette Mol and Naomi Kamoen

Learning new words in unfamiliar frames from direct and indirect teaching 121
Katharina Rohlfing, Juana Salas Poblete and Frank Joublin

The Roadsigns of Communication 131
Tania Rojas-Esponda

Incomplete Answers and the Rise-Fall-Rise Contour 140
Michael Wagner, Elise McClay and Lauren Mak

'Attention, I'm violating a maxim!' A unifying account of the final rise 150
Matthijs Westera

Multimodal Prediction of Psychological Disorder: Learning Nonverbal Commonality in Adjacency Pairs 160
Zhou Yu, Stefan Scherer, David Devault, Jon Gratch, Giota Stratou, Louis-Philippe Morency and Justine Cassell

Poster Abstracts

Gaze cue effect during language comprehension 171
Dato Abashidze, Pia Knoeferle and Maria Nella Carminati

Conversation and Incomplete Knowledge 173
Nicholas Asher and Soumya Paul

Turn-timing in naturalistic mother-child interactions: A longitudinal perspective 176
Marisa Casillas, Elma Hilbrink, Susan C. Bobb, Eve V. Clark, Merideth Gattis and Stephen C. Levinson

Putting secondary propositional meanings back on the table 179
Elena Castroviejo, Laia Mayol and E. Allyn Smith

Investigating the role of priming and alignment of perspective in dialogue 182
Simon Dobnik, Christos Koniaris and John Kelleher

Toward a Tutorial Dialogue System for Urban Navigation 185
Christopher Fox and Kristy Elizabeth Boyer

Improving Knowledge Representation to Speed up the Generation of Grammars for a Multilingual Web Assistant 188
Marta Gatius

Exploring the Role of Laughter in Multiparty Conversation 191
Emer Gilmartin, Francesca Bonin, Carl Vogel and Nick Campbell

Interpreting Instructions in a Pedestrian Routing Domain 194
Jana Götze, Johan Boye, Morgan Fredriksson and Jürgen Königsmann

"That's all. Thank you.": Emergence of Formulaic Protocols among Japanese EFL Learners . . . 197
Yasunari Harada, Mayumi Kawamura, Daisuke Yokomori and Tomoko Endo

Aboutness topic, discourse topic and the structure of discourse 199
Elena Karagjosova

Analysis and Modeling of Concern Alignment in Consensus-Building 202
Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Mika Enomoto, Yasuharu Den and Shogo Okada

Towards Structural Natural Language Formalization: Mapping Discourse to Controlled Natural Language 204
Nicholas Hubert Kirk

Abduction and parameterized semantic composition in speech-gesture integration 207
Udo Klein, Hannes Rieser, Florian Hahn and Insa Lawler

Using Learned Predictions of User Utterances to Decrease Distraction 210
Fredrik Kronlid, Staffan Larsson and Pontus Wärnestål

Making Human-Robot Quiz Dialogue More Conversational by Adding Non-Quiz Talk 213
Ivana Kruijff-Korbayova, Bernd Kiefer, Ilaria Baroni and Mattia Coti Zelati

Visual distraction test setup for an multimodal in-vehicle dialogue system 215
Staffan Larsson, Sebastian Berlin, Anders Eliasson and Fredrik Kronlid

A Corpus-Based Study of Accept and Assessment in SWBD-DAMSL 218
Yanjiao Li, Alex C. Fang and Jing Cao

Multimodal Propositions? From Semiotic to Semantic Considerations in the Case of Gestural Deictics 221
Andy Lücking

Establishing a communication system: Miscommunication drives abstraction 224
Gregory Mills

Pragmatic Markers of Common Ground in Tertiary Institution Students Sexual Discourse in Lagos State, Nigeria 226
Olawunmi Oni-Buraimoh

Toward a semantics for French short positive feedback utterances 229
Laurent Prévot

The meaning of French H*L%-contour 231
Uwe Reyle and Cristel Portes

The contribution of speech-rhythm to end-of-utterance detection 234
Franziska Schaller and Horst Müller

Successful Strategies for Ambiguity Resolution in Dialogue 237
Gesa Schole, Thora Tenbrink, Elena Andonova and Kenny Coventry

Demonstrating actions to a robot: How naïve users adapt to a robot's replication of goal and manner-oriented actions 240
Anna-Lisa Vollmer, Manuel Mühlig, Katharina J. Rohlfing, Britta Wrede and Angelo Cangelosi

Invited Talks

The Development of Reference in Early Childhood: the Roles of Communicative Motivation, Cognitive Abilities and the Caregiving Environment

Danielle Matthews
University of Sheffield
`danielle.matthews@sheffield.ac.uk`

Over the first 5 years of life children become increasingly effective communicators and remarkably aware of the co-operative nature of communication. Yet, while these developments come naturally to typically developing children, large individual differences can be observed in children's language abilities even before they begin school. I have been interested in exploring the nature of children's early communicative skills and the factors that drive development and explain individual differences. The studies I will present focus on children's ability to refer to things and to comprehend reference. I will explore how these can improve as a consequence of 1) the child's motivation to engage in and repair communicative exchanges, 2) the child's growing social and cognitive abilities and 3) the scaffolding provided by caregivers in dialogue. The emerging picture is one where children actively seek to refine their model of language, fine tuning their expectations as their experience accrues and cognitive abilities improve.

Coherence and Meaning in Situated Dialogue

Matthew Stone
Rutgers
`matthew.stone@rudgers.edu`

In face-to-face conversation, speakers use all the means at their disposal to get their ideas across. They talk, they gesture, but they also carry out practical actions in the world. These diverse actions seem to advance the communicative enterprise through common principles of discourse coherence. In this talk, I review the empirical and philosophical underpinnings of this expansive understanding of discourse coherence, and sketch a number of formal case studies analyzing situated dialogue using this approach.

Intuitions about coherence, I suggest, tap into the conventions interlocutors follow to work effectively and meaningfully with one another in conversation. These conventions establish implicit connections among communicative actions, and trigger appropriate changes to interlocutors' information and attention. Accordingly, to formalize coherence, we need representations in logical form that capture what information the speaker is committed to and what entities are at the center of attention in the discourse. Both dimensions are key to model deictic reference in situated utterances, to capture the relationship of gesture and speech, and to track how practical demonstrations update the conversational record.

This talk describes joint work with Alex Lascarides (Edinburgh) and Ernie Lepore and Una Stojnic (Rutgers).

On Variability in Pitch Accent Distributions

Marc Swerts
Tilburg University
m.g.j.swerts@tilburguniversity.edu

Speakers of germanic languages, such as Dutch and English, have been argued to use pitch accents to distinguish important from less important pieces of information in a spoken discourse, whereas listeners have been shown to be sensitive to the way such accents are distributed in an incoming utterance. For instance, when information is new or contrastive, it is typically marked with a pitch accent, and listeners find it easier to process speech when such accents indeed match the prominent information status of discourse fragments. However, when one analyses naturally produced discourse, one often observes exceptions to this general rule and variability, both between and within speakers, in how accents are distributed in spoken sentences. In this talk, I will elaborate on factors that may explain this variability, in particular focusing on the extent to which accents vary as a function of speaker type (e.g. good vs bad speakers), as a function of intonational differences between a speaker's first and second language, and as a function of the degree to which a speaker takes into account the listener's perspective on the ongoing discourse.

Full Papers

Demand Characteristics as a Tool for Evaluating the Design of Collaborative Tasks

Ed Baggs
School of Informatics
University of Edinburgh
`e.baggs@ed.ac.uk`

Abstract

It is proposed that more attention should be paid to demand characteristics in collaborative tasks. The paper focuses on joint problem-solving tasks of the type typically used in dialogue research. The impact of demand characteristics in these tasks—specifically, the presence of discrepancies between how researchers believe a task to be and how it is perceived by subjects—is often difficult to evaluate from published write-ups, because attempts to identify such confounds are typically unsystematic. This need not be the case. Methods exist to evaluate the validity of our descriptions of a given task. In addition, tasks involving dialogue have a unique feature, namely the openness of the exchange between subjects, which allows us to directly observe what kinds of cues subjects make use of while completing the task. We can exploit this openness to evaluate and improve task methodology; this last point is illustrated with some examples from the HCRC map task corpus.

1 Introduction

It is a commonplace to observe that context plays an essential part in conversation, but this is misleading. The word *context* implies that the real business of an interaction is the language used, and that everything else is mere scaffolding. From the point of view of a given individual, however, this is simply not the case. An individual is only ever trying to accomplish a *task*; the language used during a task is at best a means, not an end in itself (cf., Cohen, Levesque, Nunes, & Oviatt, 1990).[1]

On this way of seeing things, conversational transcripts, and other records of the language used during the completion of an experimental task, are traces of what happened during the completion of the task, analogous to a series of footsteps left on a beach. In much of the empirical work carried out on dialogue and interaction, the implicit goal has been to derive general truths about language use from these kinds of linguistic traces taken from experimental data and speech corpora (Schober, 2006). The ultimate goal here seems to be to come up with a general theory of communication, so we'll call this way of doing things the general theory–directed approach. The present paper adopts an alternative, task-directed approach. Here, these linguistic traces are seen as a tool for understanding the *tasks* in which linguistic data originated. In particular, this is proposed as a method for evaluating the internal validity of tasks: is our description of a task consistent with how the task is really perceived by those carrying it out?

A task here is understood in a commonsense way as any (language-involving) goal-directed phenomenon we are interested in explaining; exactly what the nature of a given task is is subject to revision following empirical investigation. What's needed is that, for a given task, we have a good way of assessing what exactly is going on when people carry it out: what specific mechanisms are employed? This is necessary if we want to know how confident we should be about our description of the task of interest, and, ultimately, about the extent to which we are justified in making general conclusions from results specific to the task. Below I propose that the concept of demand characteristics can be adapted as one tool for addressing these issues.

Demand characteristics, on the definition given below, are something common to all psychological experiments as well as to many other situations

---

[1] Couldn't the task be simply to have a conversation? Perhaps, but even then the goal is not to produce a conversational record for its own sake, but to gain knowledge from other people, or to tell them a story, to pass the time, etc.

where someone is following instructions. There are two reasons for narrowing the focus here to experiments on dialogue: 1) I believe the literature on dialogue could only benefit from more attention being paid to task demands and accompanying issues with validity, and 2) dialogue tasks produce data that is particularly useful for developing ideas about demand characteristics themselves, because the open exchange that occurs between the individuals carrying out the task can often allow researchers to reconstruct what was going on as the task was being carried out. Section 4 onwards will be concerned with the second point.

2 Demand characteristics

What are demand characteristics? The concept of demand characteristics is sometimes confused with the more specific 'good subject effect', the idea that subjects want to help the experimenter get useful results, and so behave in the way they think is expected of them. The concept is much deeper than this, however. Ultimately, it is about what tasks look like from the subject's point of view (Kihlstrom, 2002): demand characteristics are the properties of a task situation as perceived by the person carrying out the task.[2] Orne (1962), who introduced the term, wrote:

> 'The subject's performance in an experiment might almost be conceptualized as problem-solving behavior; that is, at some level he sees it as his task to ascertain the true purpose of the experiment and respond in a manner which will support the hypotheses being tested. Viewed in this light, the totality of cues which convey an experimental hypothesis to the subject become significant determinants of subjects' behavior. [...] These cues include the rumors or campus scuttlebutt about the research, the information conveyed during the original solicitation, the person of the experimenter, and the setting of the laboratory, as well as all explicit and implicit communications during the experiment proper. *A frequently overlooked, but nonetheless very significant source of cues for the subject lies in the experimental procedure itself, viewed in the light of the subject's previous knowledge and experience.* For example, if a test is given twice with some intervening treatment, even the dullest college student is aware that some change is expected, particularly if the test is in some obvious way related to the treatment.' [emphasis added]

One technique researchers have used to try to mitigate the confounding effect of subjects' expectations about an experiment is to deceive them as to the true purpose of the task. As Orne was aware, however, the efficacy of such deceptions is hard to assess from subjects' behaviour alone: a subject might appear to be behaving as the experimental manipulation predicts, but we do not necessarily know if this is a spontaneous response that reflects how the subject would behave outside of the laboratory, or if it is a more narrow response to some particular perceived cue in the set-up. And further, there exists a 'pact of ignorance' between subject and experimenter: subjects presumably have no wish for their data to be discarded from the analysis, and researchers do not wish to have to replace subjects, so it is in the interests of neither for the experimenter to probe too hard about what the subject was thinking during the task, lest the data should have to be rejected (Orne, 1969).

A note here on deception. It might be contended that this kind of deception is not relevant to tasks in the cognitive literature on language use, where everything is as it seems, and subjects are merely being asked to solve a problem set by the experimenter; in the map task, considered below, subjects are explicitly given roles as either the giver or follower of instructions, and are then simply instructed to carry out the task between themselves. We cannot assume, however, that things are so straightforward. Some of the most famous psychological experiments of the past sixty years or so—the ones our subjects are most likely to be aware of (such as the Milgram experiment)—*do* involve deception. Moreover, the undergraduate students that volunteer for the deceptive experiments are the same as those that volunteer for the non-deceptive ones. And so we must proceed on the assumption that any task that can be perceived as involving deception is likely to be so perceived. That is, even if we are not trying to deceive, we

---

[2] Following Kihlstrom, I'll continue to use the term 'subject' in preference to 'participant', as it is a more precise descriptor of the volunteer's role in the systematically designed tasks considered here.

still have to consider the possible presence of deception from the subject's point of view.

Whether a task has confounding demand characteristics or not is not simply an objective property of the task. It should be stated clearly that demand characteristics are specific to the subject, and can be located only in the interaction of the subject with the task as a whole. Demand characteristics overlap, in this sense, with James Gibson's concept of affordances (Gibson, 1979). It is tempting to suggest that demand characteristics are an instance of affordances specific to the laboratory, but this would be misleading. Affordances are *opportunities* for action, perceivable by an organism in the relation between external structure and its own ability to act upon that structure. Demand characteristics, by contrast, are contractual in quality: subjects in an experimental situation have committed themselves to carry out the task the experimenter has set for them; a response might be required even if no meaningful action is perceived (for example, a forced choice might have to be made between two stimulus items that appear the same). Different subjects will perceive a given task differently because they bring different things into the experiment: some will arrive with knowledge that's relevant to the task hypothesis: perhaps they have participated in a similar task before, or they might have had some other experience or training that makes them well-placed to detect the hypothesis. Researchers are generally aware of these problems, and try to avoid, for instance, testing the same subjects on similar tasks, or on different variants of the same task.

Despite this complication—that different cues are available to different subjects—we can still hope to identify properties within a task procedure and set-up that are likely to generate problematic demand characteristics. It may be useful to conceptualize the kinds of cues present in a given task as likely to tilt the resulting behaviour either towards or away from that predicted by the research hypothesis. I'll call these positive and negative demand characteristics, respectively (these labels are intended to be analogous to 'false positive' and 'false negative', rather than to imply good and bad). It then becomes possible to think of the (internal) validity of an experiment as a function of the cues present. This is represented schematically in Fig. 1. Note that if a task produces cues that consistently tilt behaviour one way or the other, then the task falls outside the shaded zone, and the task procedure should be considered insufficiently sensitive to detect the behaviour of interest. Note also that it is not enough for a task to fall within the shaded area for it to be considered *externally* valid—that is, a genuine result may still fail to generalize outside of the task, if the task is a poor model of the phenomenon of interest. Fig. 1 applies only to tasks that might appear to involve deception, or where the true research hypothesis is otherwise hidden from the subject; the situation may be different for non-deception tasks, such as, say, a test designed as a simple evaluation of a person's ability in some area (an IQ test is Orne's example); here, positive demand characteristics may merely serve to increase motivation.

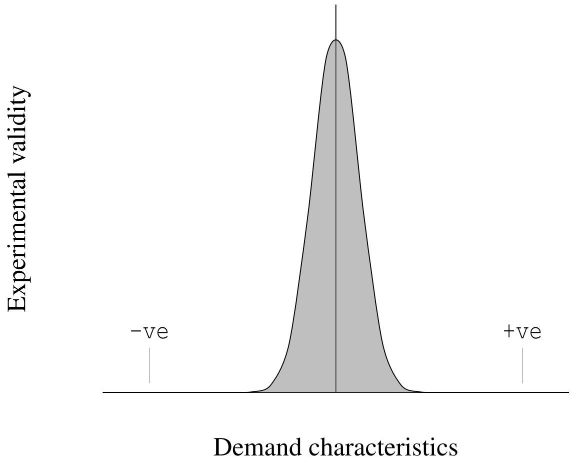


Figure 1: Schematic of the space of possible tasks (in which the research hypothesis is hidden from subjects), showing experimental validity as a function of demand characteristics; validity rapidly declines as demand characteristics push subjects' behaviour towards (positive demand characterstics) or away from (negative demand characteristics) the research hypothesis.

3 Dialogue tasks

There are a variety of ways in which researchers have attempted to study dialogue in the laboratory. I'll here consider one common class of tasks—referential communication games (Yule, 1997)—in which two subjects are recruited to complete a problem-solving task together (I'll ignore versions that use confederates). Routinely, these tasks involve constraints placed on the pairs over how they are allowed to solve the task. Often each member of the pair is given separate materials that they have private access to and the task is for one member to communicate something about the structure of their materials to the other, using only linguistic expressions.

To reiterate the point at the beginning of this paper: the approach being advocated here is concerned with explaining specific mechanisms involved in the completion of particular tasks. To be clear, by mechanisms here I do not mean internal algorithmic-level descriptions of steps involved in carrying out a task. Instead, I propose to understand a task environment, which includes oneself and other people, as providing a set of possible resources that can be assembled in pursuit of a goal (Wilson & Golonka, 2013). A mechanism, then, is a way of assembling those resources.

That being the case, why should we be interested in these referential communication games? These tasks are not interesting *per se*; they exist because they were devised to advance some general theory about how communication works, not because the researchers who devised the tasks had some inherent interest in this kind of game (for example, early versions of these games explicitly instantiated an information theoretic code model of language as a signal transmitted between an encoder and a decoder; the tasks were employed as a means of disrupting feedback; see Krauss & Weinheimer, 1966). The answer is that we don't currently have a well-developed way of going about the study of collaborative activity that primarily seeks to explain tasks; we do, however, have corpora from existing tasks, such as the map task, below, that can be used as immediate material for developing such an approach. So the following is a preliminary attempt to develop the tools of a task-directed approach, drawing on an existing corpus of data.

4 Demand characteristics in the map task

The HCRC map task (Anderson et al., 1991) is an interesting case in terms of demand characteristics because it was set up not to test a single hypothesis, but to test several hypotheses at once, and to produce a corpus of data that could be used to investigate an open-ended set of research questions. Meanwhile, the concept of demand characteristics, as defined, is only meaningful relative to a single, specific research hypothesis. One might think, then, that the concept would be hard to apply here. Nonetheless, it's easy to identify cues that people are aware of while carrying out the task, and we can talk about these cues in general terms; we can do this by examining the recordings and transcripts from the corpus (available at http://groups.inf.ed.ac.uk/maptask/). Note that the following is not meant to be a discourse analytic discussion of the task. Looking for demand characteristics should be seen as part of the experimental design and evaluation process; it is a way of asking whether our description of the task matches the reality from the subject's point of view. The discussion of the map task here is meant to demonstrate that this can in principle be achieved, in part, by examining the open exchange that goes on as people carry out the task.

In this task, an instruction giver sits in front of a map with a predefined route drawn on; the goal is to communicate this route to an instruction follower who can't see the instructor's map, and for the follower to reproduce that route on their own map. Subjects were told this goal explicitly: 'Subjects were told that the goal of the task was to enable the Giver's route to be drawn on the Follower's map, that the Giver's and Follower's maps might be different in some respects, and that both participants could say whatever was necessary to complete the task, but that neither could use gestures.'

Examples of the maps can be seen in Figures 2 and 3. The instructor had the map on the left; the follower's completed map is shown on the right. I'll here look at three exchanges that illustrate some effects of demand characteristics in this task.

The first exchange (from a pair coded as q1nc2 in the corpus) I present as evidence that the constraints on communication described in the instructions given to subjects are only partly true as a description of what actually happened in the task. Specifically, the rule that 'neither could use gestures' can only have been partly followed (g is the instruction giver, f the follower; I have added the comment and punctuation):

> g — and you should be kind of ehm two and a half inches away from the right-hand side of the page just now
> f — oh [uhh...] no
> g — no
> g — where are you?
> f — my inches must be different from

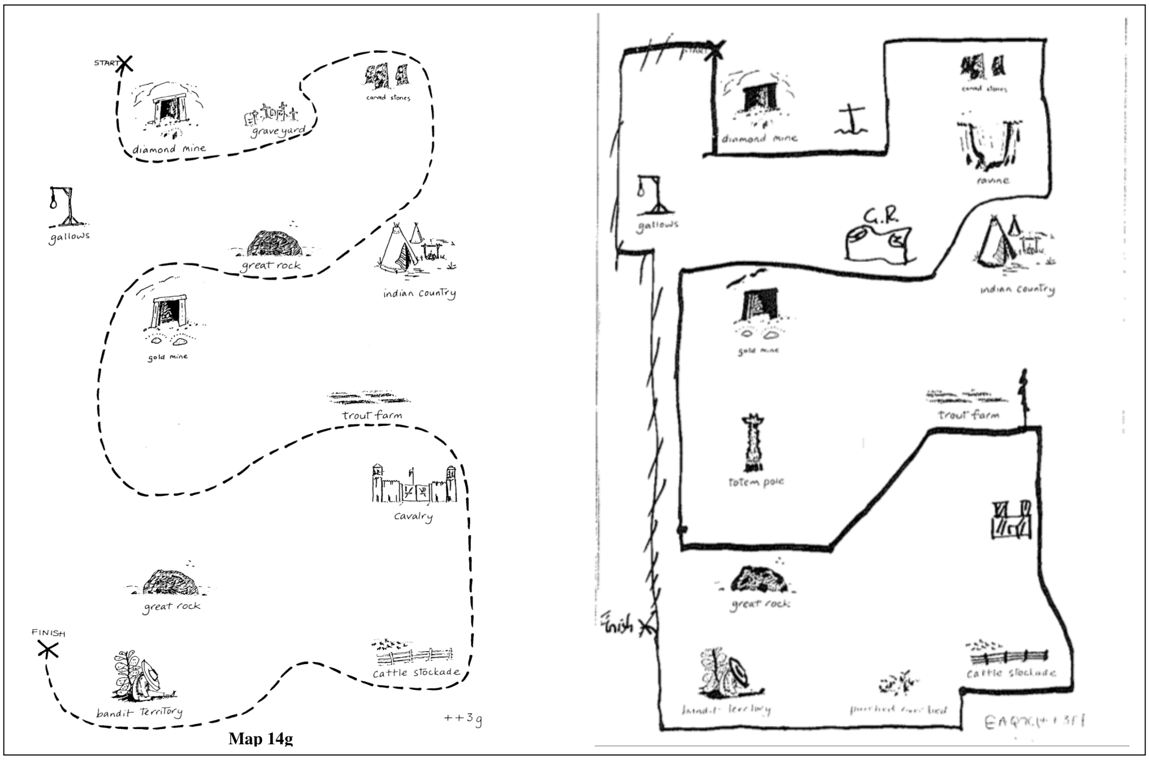


Figure 2: Map task conversation q7ec1—the route giver's map is on the left, the completed follower's map on the right; in this trial, the pair could make eye-contact; they were both male, and knew each other beforehand; recording duration 5'58" (map images are copyright Human Communications Research Centre 2007, and are available under a creative commons licence, cc-by-nc-sa)

> yours 'cause I'm not even halfway across the page
> f — I should be away at the other s– side of the page?
> g — you should be kind of at the right-hand side
> f — how l– how big's your page?
> g — er
> f — is it that size? [f shows the back of her map to g]
> g — uh-huh
> f — uh-huh

This exchange in fact comes from a no-eye-contact trial, in which there was a barrier between the pair. The follower can be heard on the recording wielding the page. What's not seen in the transcript is that the instructor breathes in, perhaps apprehensive about what has just happened, as if she is worried that they have just broken the rules and so will have to be ejected from the experiment. Of course, by normal standards, this is a perfectly sensible thing to do: showing something to someone to confirm that you're both talking about the same thing. (Even more sensible would be for the instructor to pass her map over the barrier for the follower to copy out the route directly. None of the participants did this, of course; they would have been ejected.) Here, then, is one instance of gesturing that found its way into the corpus. Video recordings of the sessions (not available online) no doubt contain countless other instances, particularly if we consider facial expressions as gestures.

The lesson here is perhaps that if you want your subjects to behave towards one another in a specific way, it is not reasonable to place the burden of maintaining that behaviour on the subjects themselves. The subjects did not have visual access to each other's maps. This was more or less guaranteed by the layout of the furniture in the laboratory. They did, however, have continual access to each other's gestures, and to their own ability to produce gestures. Given how ubiquitous gesturing is in life outside the laboratory, it would seem to require considerable effort to deliberately suppress this behaviour.

The second exchange is from the pair whose maps are shown in Fig. 2. This exchange con-

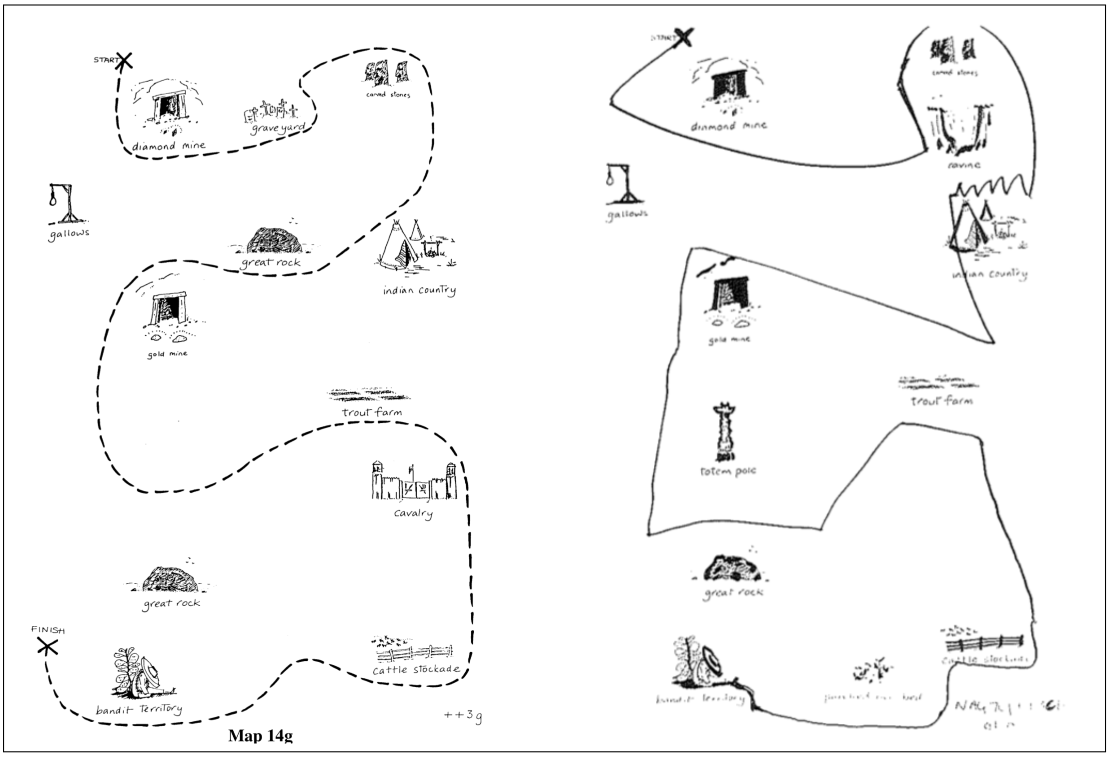


Figure 3: conversation q7nc1—no eye-contact, both female speakers, knew each other beforehand, recording duration 6'44"

tained a false finish, hence the crossed-out line on the left. The completed follower's map also features some extra landmarks, which the instructor insisted be drawn in (the initial maps differed in the placement of some landmarks). These both suggest that the pair were motivated to perform the task well.

The recording of this pair also reveals another aspect of dialogue tasks which is absent from non-dialogue tasks. It's clear from the recording of this exchange (though again, not the transcript) that the instructor is trying to make the follower laugh as they complete the task. He repeatedly instructs the follower to 'hang a left', instead of the more mundane 'turn left', and does so with audible delight. At the point above the Indian country on the right hand side:

> g — until you get to the indian country
> then you do a wee chicane
> g — turn left above the indian country

Between these two utterances the follower can be heard chuckling. It seems fair to say that instructor is willing to sacrifice some precision here in favour of making the task more enjoyable. Here is a demand characteristic peculiar to tasks that allow interaction: a joint task is also a social activity between subjects. Whether this is something to be concerned about will depend on the research question we are interested in answering.

Finally, look at Fig. 3. This is the same map as in Fig. 2, completed by a different pair. There is a salient feature on the instructor's map towards the top, where the route makes an 'S' curve around the graveyard. The instructor in Fig. 3 draws this to the follower's attention and tells the follower to go 'back towards the right' (this pair started at the finish point, hence 'right' and not 'left'). This bend can be seen on the follower's map in Fig. 3. However, none of the other completed versions of this map (each map was completed by eight different pairs) features this curve. The goal of the task as interpreted by the pairs seems to have been to avoid hitting the landmarks. It is worth emphasizing this because it conflicts with the assumption that the goal defined by the instructions—'to enable the Giver's route to be drawn on the Follower's map'—is well defined. Anderson et al. assume it is, and that this allows for an objective measure of communicative success: 'Because the correct solution to the problem is well defined, *successful communication can be measured* in terms of the extent to which the achieved route corresponds to the model.' [emphasis in original] If people are partly using a landmark-oriented strategy, then the standard measure of success (absolute deviation from the path) is strictly measuring a different thing from what subjects are alert to: it measures whether the path is in the right place in absolute terms, not whether it is in the right place relative to the landmarks.

In summary, these exchanges provide evidence for three properties of the task not acknowledged in the original description in Anderson et al. (1991): 1) The instruction that subjects cannot use gestures creates an artificial burden on subjects to monitor their own behaviour. 2) The task has properties not present in individual problem-solving tasks: participants here are sometimes attempting to amuse each other; this may introduce a discrepancy between the overall goal of the task as the subjects see it and the task as the researchers assume it to be. And 3) the route, as interpreted by most pairs in the map task, is landmark-oriented, and not absolute, as assumed by the researchers. This partially undermines the claim that the task has an objective measure of success. In general, we might want to consider that objective measures of communicative success are a fiction; communicative success can only be defined relative to the goal from the point of view of whoever is trying to accomplish it. Any research question that hinges on communicative success should be alert to such discrepancies between the thing measured and the tool used to measure it. Indeed, anyone using task corpus data to investigate a specific research question should try to evaluate the demand characteristics of the task relative to that question. These three observations can all be used to make better sense of the behaviour in this particular task.

To repeat, the purpose of this discussion is to demonstrate how we can take advantage of the open exchange of dialogue to evaluate the suitability of an experimental methodology for addressing a given research question, and to improve that methodology in subsequent versions of the task. This evaluation can be done in a rigorous way: produce a description of the task goal, then look for counter-evidence that that's what the goal is from the subject's point of view; describe your dependent measures, then look for counter-evidence that these are measuring what you think they are measuring; and so forth. To be sure, this is not guaranteed to detect every possible confound, but it can surely detect some.

5 Detecting demand characteristics

What we are interested in here, is detecting demand characteristics in situations where the cues are not well understood and where unknown confounding cues may be present. Orne (1969) described three main methods for doing this. He called these methods 'quasi-controls'. All of his methods seek to recruit the subject as a co-investigator. Orne was interested in hypnosis; he developed the concept of demand characteristics in order to ask questions such as this: are hypnotized subjects really under the control of the hypnotist, or might they merely be behaving in the manner they think they're expected to, because of the peculiarities of the situation? The techniques may be partly applicable to dialogue research too.

The first method is simple post-test inquiry: ask the subject what they thought they were doing. Such inquiries are presumably widely conducted nowadays, but are less commonly reported. It is not clear why this should be the case. These questionnaires are in part suspect, of course, because of the pact of ignorance mentioned above: research participants do not wish to be ejected from the analysis, and so, if they did in fact suspect some deception, they have an incentive to keep this to themselves. But this would still yield a set of responses consistent with the deception being valid, and even this kind of thing is not widely reported. One reason why researchers may omit the questionnaire data from the write-up is that it's seen as too difficult to summarize. If this is the case, though, then this too should be reported: if subjects do not in fact have a common idea of what it is they are doing, this may undermine an unstated assumption of the researchers, who presumably intend the task to be perceived in a uniform way. More diligent reporting of the kinds of things people say after a task should be encouraged.

Orne's second quasi-control method he called the 'non-experiment'. Here, subjects are shown the materials and the set-up, but not actually asked to carry out the task. Instead, they are asked to guess how others would respond if given these materials and asked to complete the task. This method may be of potential use in dialogue research. A possible shortcoming is that dialogues are unpredictable from the standpoint of any one participant, and perhaps even to a pair of non-participants: each member of a pair has only a partial perspective on what the task is. A pair might only be able to work out how they would perform a task by actually doing the task. Similarly, Orne noted that the non-experiment cannot be sensitive to cues that subjects themselves are not consciously aware of. Still, the method could allow researchers to see what kinds of approaches people are inclined to take going into a task.

The third method—simulation—is perhaps more specific to the kinds of question Orne was interested in. Here, 'simulating' subjects are recruited and asked to behave as if they are real subjects, that is, they're told they're in an experiment involving hypnosis, and are asked to behave as if they are actually hypnotized; there is an experimenter who is blind to who are the simulators and who are the 'reals'; the simulators' task is to make the experimenter believe they are genuinely hypnotized. It's harder to imagine where this simulation method could be applied in dialogue research.

The discussion in this paper suggests a fourth method. Dialogues have an inherent feature that allows the researcher to look in and infer directly what kinds of demand characteristics people are sensitive to: dialogues are open, in the sense that they consist of behaviours that can be observed from the outside, rather than solely of internal mental behaviour that has to be inferred by proxy. The openness of a dialogue means that subjects can be used as their own quasi-controls. The brief discussion of the map task above is intended to demonstrate the plausibility of this method. Granted, this method involves some uncertainty; it depends on inference on the researcher's part: the researcher is looking for counter-evidence that the task is perceived by the subject in the manner intended. But the method is valuable if it allows us to detect at least some potential confounds that we would otherwise be ignorant of.

It may be useful here to say something about how, specifically, one should go about attempting to detect demand characteristics for a given set of data. First, it must be reiterated that demand characteristics are not a property of the task, but of how the task is perceived by an individual subject, relative to a research hypothesis. Specifying the hypothesis is a prerequisite before you can look for potential confounds. In general it is not possible to be very precise about exactly what to look for: this will depend on the nature of the hypothesis under consideration. But we can say, in terms of the schematic depicted in Fig. 1., that in order for a study to be valid, the task should produce demands that fall in the neutral space in the middle. That is, there should ideally be nothing about the task set-up itself that is misleadingly pulling behaviour either towards, or away from, the behaviour predicted by the research hypothesis. Non-neutral demand characteristics are a threat to validity: they cast doubt on our ability to attribute behaviour to something about the psychology of the individual subject; and raise the possibility that that behaviour should in fact be attributed to the task set-up. A write-up of the study should then seek to provide the following:

1. a clear statement about what the researchers believe constitute neutral conditions for the task under investigation
2. details of attempts to establish that neutral conditions did in fact prevail for the subjects engaged in the task, and
3. details of potential confounds which the researchers were unable to rule out from the available data.

These steps should be seen as a valuable part of the experimental design and evaluation process.

Finally, it must be admitted that these proposals are not especially novel. Some published studies on dialogue do make use of some of these methods. In particular, I'll note that in Schober and Clark (1989)—a study of how well over-hearers to a referential communication game are able to make sense of a discussion they're not part of—the authors include substantial discussion, under the heading 'Subjective commentary', of both questionnaire data (for experiment two), as well as inferences drawn from analysis of the recordings (for experiment one); that is, they made use of both inquiry and openness to evaluate the experimental design. For someone reading this paper with an eye to how the task looked form the subject's point of view, these discussions are extremely useful.

6 Implications for future work

At the beginning of this paper I made a distinction between general theory–oriented and task-oriented approaches to the study of language use. The discussion about demand characteristics in the map task has arguably been consistent with either approach. I believe, however, that it is worth trying to pursue an alternative task-directed methodology that makes a strong claim to distinguish itself from the general theory–directed programme. The strong claim is this: the practice of producing language corpora from tasks as a method of studying 'dialogue' is misguided; coprus data can *only* be used as a means of evaluating the task. The reasoning here is as follows. If we want to draw general conclusion from observing specific tasks, then we need to be confident both that our description of the task is correct, and that the task itself is representative of the phenomenon we wish to model. In the case of dialogue, and the tasks used to model it, neither of these is necessarily true. Indeed, it's not clear what the scope of 'dialogue' is at all. It is clear, however, that we cannot judge *what* a task is representative of until we have a good understanding of the task itself; we have to know where to position the task on Fig. 1. One way of doing this is by appeal to quasi-control techniques for discovering demand characteristics.

A possible implication here is that the goal of a psychology of language use should not be to produce a general theory of communication at all; the goal should instead be to identify the mechanisms involved in the completion of specific tasks. This might appear a pessimistic conclusion. But it can perhaps be argued that a more modest scope has the potential to produce more tractable research questions than those commonly asked at present, and may be the only way to carry out a genuinely incremental psychology of language use.

7 Acknowledgments

This work is supported by an EPSRC grant, and by funding from the European Commission through the JAMES project (FP7-ICT-270435); I thank Holly Branigan and Jon Oberlander for their comments and discussion, and the reviewers for their useful remarks.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and speech*, *34*(4), 351–366.

Cohen, P. R., Levesque, H. J., Nunes, J. H., & Oviatt, S. L. (1990, November). Task-oriented dialogue as a consequence of joint activity. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence.* Nagoya, Japan.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton-Mifflin.

Kihlstrom, J. (2002). Demand characteristics in the laboratory and the clinic: Conversations and collaborations with subjects and patients. *Prevention & Treatment*, *5*(1).

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343–346.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783.

Orne, M. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 143–179). Academic Press.

Schober, M. (2006). Dialogue and interaction. *Encyclopedia of language and linguistics*, *2*, 564–571.

Schober, M., & Clark, H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232.

Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in psychology*, *4*(58).

Yule, G. (1997). *Referential communication tasks.* Mahwah, NJ: Lawrence Erlbaum Associates.

Update conditions and intensionality in a type-theoretic approach to dialogue semantics

Robin Cooper
University of Gothenburg
`cooper@ling.gu.se`

Abstract

We consider the interaction of update conditions for dialogue gameboards, compositional semantics and intensionality. We will concentrate on the update conditions associated with proper names and definite descriptions. It is well-known from the literature that proper names require the dialogue partner being addressed to be able to identify an individual with the appropriate name or at least a role for an individual of that name in the content of the dialogue. Slightly more controversially we will take standard uses of definite descriptions to require the dialogue partner to be able to identify (a role for) a unique individual of that description. A puzzling example from this perspective is:

(opening presents on Christmas morning – A and B have failed to get a trainset for Sam)

A: Sam is looking for the trainset
B: What trainset?
A: The one he was promised for Christmas

We will present an analysis of this in which B is required to accommodate a *type* of a situation in which there is a unique trainset.

1 Introduction

In classical formal semantics (Montague, 1973; Montague, 1974) proper names are treated as denoting the set of properties of a unique individual and singular definite descriptions are given a Russellian analysis. Neither of these analyses introduce any kind of presupposition or familiarity requirement. In dynamic semantics (Heim, 1982; Kamp, 1981; Kamp and Reyle, 1993) linguistic content is viewed in terms of update potential and familiarity constraints are introduced in respect of definite and indefinite noun-phrases. However, this work was discourse oriented and did not take into account the updating of individual dialogue participants' gameboards as in the work of Ginzburg (1994; 2012). For recent discussions of alternatives presented in the voluminous literature on the semantic treatment of singular definite descriptions see Elbourne (2012) and Coppock and Beaver (2012).

In this paper we will adopt more or less the approach of Ginzburg (2012) but try to combine it with the rigorous approach to compositional semantics introduced by Montague (1973). In the process we will show that we can treat a kind of intensionality that arises in dialogical exchange that we believe has not been treated previously in the literature on dialogue semantics. It seems also to be a kind of intensional construction which potentially poses challenges for current treatments of definiteness in general in compositional semantics, though I am not yet in a position to evaluate which current proposals might successfully compete with the proposal here. The main aim of this paper is to get a closer connection between dialogue semantics and *some* kind of compositional semantics applied to a traditional semantic concern.

The basic data we wish to account for concerns (somewhat modestly in terms of modern semantics) proper names and singular definite descriptions. If A says (1) to B, then B is required to have a gameboard which somehow identifies an individual named Sam before the content of (1) can be integrated into B's gameboard.

(1) Sam left

If B's gameboard does not provide such an individual then some kind of accommodation has to take place. We will try to say something about the nature of the accommodation processes which might be involved, including one where B is not acquainted with an individual named Sam but relies on the fact that A has identified such an individual. Similarly, if A says (2) to B, then B's gameboard must contain information which enables him to identify a unique dog before the content of (2) can be integrated into his gameboard.

(2) The dog left

If the gameboard does not provide such an individual then some kind of accommodation has to take place, including one possibility where B is not himself able to identify an appropriate individual but relies on A being able to do so. We will take a rather conservative approach to definite descriptions, using a variant of Montague's (1973) Russellian approach combined with a notion of resource situation (Barwise and Perry, 1983; Cooper, 1996) in which there is a unique individual which falls under the description.

This choice plays a role when we consider the analysis of examples involving intensional constructions. Consider (3) which is a constructed dialogue based on a non-dialogical example presented by Max Cresswell.

(3) *(opening presents on Christmas morning – A and B have failed to get a trainset for Sam)*
A: Sam is looking for the trainset
B: What trainset?
A: The one he was promised for Christmas

The intended reading for A's first utterance is a *de dicto* one where the definite description *the trainset* is within the scope of the intensional verb *look for*.[1] There is no trainset under the Christmas tree. Both A and B know this and one senses a drama about to unfold. As one might expect on such a reading there is no requirement that B be able to identify a unique trainset on the basis of his gameboard. Furthermore, this is distinct from the non-intensional cases above where B had the option of relying on A being able to identify the appropriate trainset. There simply is no trainset which Sam is looking for. That, one suspects, is the point of A's initial remark. B knows there is no trainset. It is not the case that B misinterprets A's assertion as *de re*, that is, as referring to some particular physically existing trainset. That certainly would be a possible interpretation in a different context. But here, we assume, the background to this dialogue could be that A and B promised Sam a trainset for Christmas and agreed that B should buy it. B has subsequently forgotten this promise and knows that no trainset has been bought. Thus the clarification request is not a request for a reference to any particular physically existing trainset, but rather a request for an explanation of why a trainset is expected to be under the tree. This is potentially a problem for previous treatments of clarification such as Ginzburg and Cooper (2004), Purver and Ginzburg (2004), Ginzburg (2012) and Cooper (2013) where clarification is treated in terms of providing values for referential parameters.

If we think of the prior gameboard requirements engendered by utterances as being like presuppositions then it seems natural for the embedding of a noun-phrase in the scope of an intensional verb to block their projection to the root of a sentence in a compositional semantics. But there is, of course, a problem with this, as shown by (3). If the requirement that there is a unique trainset is blocked compositionally by the intensional verb how is it that B can ask his clarification question and A can give her answer, apparently referring to a trainset? Our proposed solution to this will treat the intensional verb as a *filter* rather than a *plug* (Karttunen, 1973). Our analysis will exploit the fact that we are using type theory in the manner proposed by Ginzburg (2012) and Cooper (2012). We shall propose that what gets passed up is not the requirement that the gameboard identifies a unique trainset but that an appropriate *type* of situation where there is a unique trainset is available on the gameboard. It will be important for our analysis that we are dealing not with traditional presuppositions but with constraints on gameboards. Whereas it may be trivial to claim that a given type exists, it is another matter altogether to require that an agent has such a type available on their gameboard.

The "type of situations where there is a unique trainset" is not of itself a very informative type. We can understand why B may ask a clarification question. We will suggest that the effect of B's question is to ask for a subtype of this type in

---

[1] The alternative is a *de re* reading where *the trainset* has wide scope outside of the scope of *look for*. This would, for example, be appropriate for a situation where there is a particular trainset under the Christmas tree and Sam is looking for the appropriate parcel.

which more information is given about the trainset involved in a situation of this type. A's response to the clarification request is a noun-phrase similar to examples discussed by Hulsey and Sauerland (2006) and Grosu and Krifka (2007). One may think that it is ambiguous between an extensional reading where it refers to a particular existing trainset and a "reconstructed" reading where the semantic contribution of the head is embedded below *promised*. We will claim below that despite the fact that there is no trainset it is the extensional reading that is relevant here and that the intensionality derives from an update process akin to modal subordination (Roberts, 1987). In this case there is no modal and we will call the process *type subordination*. The idea is that the clarification response is used to update a "subordinate" type introduced within the type representing the commitments (or FACTS) on the gameboard. In this case the updated type will be the "type of situations containing a unique trainset which was promised to Sam for Christmas".

2 Proper names

We will follow Ginzburg (2012) in using TTR (Type Theory with Records) (Cooper, 2012) to model both dialogue gameboards and compositional semantics. For orientation, we will first show how to recapture something very close to Montague's (1973) original treatment of proper names within TTR. We will then show how this can be modified into a semantics introducing update conditions.

Intransitive verbs like *leave* have as their content functions which map records containing an individual to a type of situations where that individual leaves.[2] (4a) is the function which is the content of *leave* and (4b) is the type to which the content of intransitive verbs are required to belong. Hence (4a) is of the type (4b).

(4) a. $\lambda r{:}\big[\text{x:}\mathit{Ind}\big].\ \big[\text{e:leave}(r.\text{x})\big]$[3]

b. $(\big[\text{x:}\mathit{Ind}\big]{\rightarrow}\mathit{RecType})$

c. *Ppty* – "property"

We abbreviate the type (4b) as (4c), that is the type of properties. Properties map a record containing an individual in a field labelled 'x' to a record type containing a type of situation. Record types serve as propositions. They are "true" if there is a situation of the type and false otherwise. The type *Ppty* corresponds to the type $\langle e, t\rangle$ in Montague semantics, mapping individuals to truth-values except we map to a type corresponding to a "proposition" so we are closer to the type $\langle e, p\rangle$, functions from individuals to propositions, introduced by Thomason (1980) and work in property theory as in for example Fox and Lappin (2005).

Montague's (1973) treatment of proper names was to treat them as functions from properties of individuals to truth values. To mimic this treatment we treat them as functions from properties to record types (corresponding to types of situations or "propositions"). That is, functions of the type (5a).

(5) a. $(\mathit{Ppty}{\rightarrow}\mathit{RecType})$

b. *Quant*

c. $\lambda P{:}\mathit{Ppty}.\ P(\big[\text{x=sam}\big])$

The notation r.x refers to the object in the x-field in the record r. We abbreviate (5a) as (5b) indicating that we are following Montague in treating noun-phrases as (generalized) quantifiers (Barwise and Cooper, 1981). (5c) is our reconstruction of Montague's basic treatment of the noun-phrase *Sam* where we use 'sam' to represent a particular individual.

Interpreting the sentence *Sam left* involves applying the function (5c) to (4a) which (after two applications of β-reduction[4]) returns (6), that is, a type of situations where Sam left.

(6) $\big[\text{e:leave(sam)}\big]$

We shall address two problems with this basic treatment of proper names: (i) it does not account for the fact that a proper name can refer to different individuals, an important source of misunderstanding which we wish to be able to analyze in dialogue semantics (ii) it does not give us any way of placing the requirement on the interlocutor's gameboard that there already be a person named Sam available in order to integrate the new information onto the gameboard. As Ginzburg (2012) points out, the successful use of a proper name

---

[2] We will not treat tense here.

[3] In contrast to Cooper (2012) and elsewhere, we will use the dot-notation for λ-abstraction. We previously would have represented this function as: $\lambda r{:}\big[\text{x:}\mathit{Ind}\big]\ (\big[\text{e:leave}(r.\text{x})\big])$.

[4] See Cooper (2012), section 2.8 for a presentation of the rather special nature of function application in TTR.

to refer to an individual a requires that the name be publically known as a name for a. We shall address both of these problems by making the interpretation of the proper name be a function that maps a context – that is, a situation modelled as a record (Ginzburg, 2012) – to a quantifier. That is, a function of type (7a).

(7) a. $(\mathit{Rec}{\rightarrow}\mathit{Quant})$

b. $\lambda r{:}\begin{bmatrix}\text{x}{:}\mathit{Ind}\\ \text{e}{:}\text{named(x, “Sam”)}\end{bmatrix}.$
$\lambda P{:}\mathit{Ppty}.\ P(r)$

The basic idea is that this function (7b) can be used to update a context of the type specified for the first argument of the function, i.e. a context where there is an individual named Sam. We will change all interpretations to be such “update functions”. In this paper we are not interested in specifying what requirements the intransitive verb *leave* may place on the context so we will let it be defined on any context (that is, any record) and will define it to be a function that returns (4a) no matter what the context is. The new function is given in (8).

(8) $\lambda r_1{:}\mathit{Rec}.\lambda r_2{:}\begin{bmatrix}\text{x}{:}\mathit{Ind}\end{bmatrix}.\ \begin{bmatrix}\text{e}{:}\text{leave}(r_2.\text{x})\end{bmatrix}$

We use (9a) as the general schema of functions which combine the interpretations of two constituents α and β.[5]

(9) a. $\lambda z.\alpha(z)(\beta(z))$

b. if $\alpha : (T_1 \rightarrow (T_2 \rightarrow T_3))$ and
$\beta : (T_4 \rightarrow T_2)$
then the *combination of* α *and* β *based on functional application* is
$\lambda r{:}\begin{bmatrix}\text{f}{:}{T_1}^{\text{f.}}\\ \text{a}{:}{T_4}^{\text{a.}}\end{bmatrix}.\ \alpha(r.\text{f})(\beta(r.\text{a}))$

c. $\lambda r{:}\begin{bmatrix}\text{f}{:}\begin{bmatrix}\text{x}{:}\mathit{Ind}\\ \text{e}{:}\text{named(f.x, “Sam”)}\end{bmatrix}\\ \text{a}{:}\mathit{Rec}\end{bmatrix}.$
$\begin{bmatrix}\text{e}{:}\text{leave}(r.\text{f.x})\end{bmatrix}$

(9b) is the combination rule we use. Note that the types T_1 and T_4 represent the restrictions on the context associated with α and β respectively and that both these restrictions are passed up to the combined interpretation, though embedded under the additional labels ‘f’ and ‘a’ respectively (mnemonics for “function” and “argument”). The reason for the addition of these labels is to avoid any unwanted label clash if T_1 and T_4 should happen to contain the same label. The notation $T^{\pi.}$ where π is a path (a sequence of labels) means a type like T except that any path that occurs as an argument to a predicate is prefixed by π. (9c) is the result of combining (7b) and (8) using (9b), after β-reduction.

How do we use (9c) to place constraints on the interlocutor’s gameboard? The idea is that the domain type in (9c) should be used to place a requirement on what is already present in the gameboard. The part of the gameboard that is relevant is that which represents the agent’s view of what has been established in the dialogue so far, that is the field which is labelled FACTS in Ginzburg (2012) and commitments in Larsson (2002) and Larsson and Traum (2001) and other work in the computational information state approach based on Ginzburg’s gameboard theory. Both Ginzburg and Larsson regard this field as containing a set of propositions. Ginzburg (2012) furthermore regards the propositions as being Austinian, that is, records each with a field for a situation and a type. What we shall use for update here, however, is a single record type which is used to keep track of the collected content of the dialogue. It seems much easier to understand how to use the kind of update functions discussed above to update this type. It corresponds to proposals within DRT for using a single DRS to keep track of the contribution of a discourse and to express anaphoric relations across sentences in a discourse. This is not meant as an argument against using Austinian propositions which seem independently useful. Perhaps the gameboard needs to contain a set, sequence or string of Austinian propositions in addition to the kind of type that we are talking about. It seems that our record type could be derived from a string or sequence of propositions representing the history of propositional updates to the gameboard by merging all the types in the Austinian propositions into a single large type which represents the commitment of the dialogue to the existence of a situation of that type. We will, however, not pursue this further in this paper.

---

[5] This is the λ-calculus version of the S-combinator in combinatorial logic. It is the standard manoeuvre for combining meanings, that is functions from context to contents, in compositional semantics, where the content of a phrase is the result of applying the content of one constituent to the content of the other constituent.

(10) a. $T_{i+1} = \left[\text{pr}{:}T_i^{\text{pr.}}\right] \underset{.}{\wedge} T_j$

b. *A boy hugged a dog. A girl stroked a cat*

c.
$$\left[\begin{array}{lcl}
\text{pr} & : & \left[\begin{array}{l}
\text{pr}{:}\mathit{Rec}\\
\text{x}{:}\mathit{Ind}\\
\text{c}_{\text{boy}}{:}\text{boy(pr.x)}\\
\text{y}{:}\mathit{Ind}\\
\text{c}_{\text{dog}}{:}\text{dog(pr.y)}\\
\text{e}{:}\text{hug(pr.x,pr.y)}
\end{array}\right]\\
\text{x} & : & \mathit{Ind}\\
\text{c}_{\text{girl}} & : & \text{girl(x)}\\
\text{y} & : & \mathit{Ind}\\
\text{c}_{\text{cat}} & : & \text{cat(y)}\\
\text{e} & : & \text{stroke(x,y)}
\end{array}\right]$$

We will thus assume that the commitments of the dialogue are kept track of in a field on the gameboard which contains a single type. Initially, before any commitments have been made in the dialogue, this type will be *Rec*, the type of records. The basic rule for updating a commitments type T_i with a new type T_j to obtain the current commitments type T_{i+1} is given in (10a). The label 'pr' ("previous") is used to ensure that label clash does not occur and it also gives us a way of maintaining a record of the order in which various commitments were introduced. Previous contributions become more and more deeply embedded as the dialogue progresses. This is represented by (10c) which shows one possible way of representing the commitments of the discourse (10b). The boy and the dog are held distinct from the girl and the cat despite the fact that the labels 'x' and 'y' have been reused. The symbol $\underset{.}{\wedge}$ in (10a) represents the merge operation on types as discussed in Cooper (2012). In the simplest case for record types which do not share any labels this involves forming a type with the union of the two sets of fields from the types being merged. (11) gives a hint of the general strategy for treating anaphora in such a system, although that is not the subject of the present paper.

(11) a. *A boy hugged a dog. He stroked a cat*

b.
$$\left[\begin{array}{lcl}
\text{pr} & : & \left[\begin{array}{l}
\text{pr}{:}\mathit{Rec}\\
\text{x}{:}\mathit{Ind}\\
\text{c}_{\text{boy}}{:}\text{boy(pr.x)}\\
\text{y}{:}\mathit{Ind}\\
\text{c}_{\text{dog}}{:}\text{dog(pr.y)}\\
\text{e}{:}\text{hug(pr.x,pr.y)}
\end{array}\right]\\
\text{x=pr.x} & : & \mathit{Ind}\\
\text{c}_{\text{male}} & : & \text{male(x)}\\
\text{y} & : & \mathit{Ind}\\
\text{c}_{\text{cat}} & : & \text{cat(y)}\\
\text{e} & : & \text{stroke(x,y)}
\end{array}\right]$$

Here we use a manifest field $\left[\text{x=pr.x}{:}\mathit{Ind}\right]$ (Cooper, 2012) which requires that the individual in the x-field is identical to the individual in the pr.x-field.

The strategy of updating types in this way to model growing numbers of commitments as the dialogue progresses is essentially similar to using a DRS to keep track of commitments. Types can, among other things, model DRSs and our use of types in modelling gameboards might be seen as related to the psychological perspective on DRT presented by Zeevat (1989). Thinking of commitments in terms of a type which grows during the course of a dialogue is also closely related to Stalnaker's (1978; 2002) notion of common ground. Instead of thinking of an agent's view of the common ground as being a set of possible worlds which gets smaller as the dialogue progresses we think of it as a type of situation which gets more refined and thus places more restrictions on the nature of the situation corresponding to the commitments.

Suppose that the commitments type on the gameboard is (12a). According to (9c), we are wanting to match (12a) with the type (12b).

(12)

a. $$\left[\begin{array}{l} \text{pr}: \left[\begin{array}{l} \text{pr}:\mathit{Rec}\\ \text{x}:\mathit{Ind}\\ \text{c}_{\text{boy}}:\text{boy(pr.x)}\\ \text{c}_{\text{named}}:\text{named(pr.x, “Sam”)}\\ \text{y}:\mathit{Ind}\\ \text{c}_{\text{dog}}:\text{dog(pr.y)}\\ \text{e}:\text{hug(pr.x,pr.y)} \end{array}\right]\\ \text{x}:\mathit{Ind}\\ \text{c}_{\text{girl}}:\text{girl(x)}\\ \text{y}:\mathit{Ind}\\ \text{c}_{\text{cat}}:\text{cat(y)}\\ \text{e}:\text{stroke(x,y)} \end{array}\right]$$

b. $$\left[\begin{array}{l} \text{f}: \left[\begin{array}{l} \text{x}:\mathit{Ind}\\ \text{e}:\text{named(f.x, “Sam”)} \end{array}\right]\\ \text{a}:\mathit{Rec} \end{array}\right]$$

c. $$\left[\begin{array}{lcl} \text{pr.pr} & : & \mathit{Rec}\\ \text{pr.x} & : & \mathit{Ind}\\ \text{pr.c}_{\text{boy}} & : & \text{boy(pr.x)}\\ \text{pr.c}_{\text{named}} & : & \text{named(pr.x, “Sam”)}\\ \text{pr.y} & : & \mathit{Ind}\\ \text{pr.c}_{\text{dog}} & : & \text{dog(pr.y)}\\ \text{pr.e} & : & \text{hug(pr.x,pr.y)}\\ \text{x} & : & \mathit{Ind}\\ \text{c}_{\text{girl}} & : & \text{girl(x)}\\ \text{y} & : & \mathit{Ind}\\ \text{c}_{\text{cat}} & : & \text{cat(y)}\\ \text{e} & : & \text{stroke(x,y)} \end{array}\right]$$

d. $$\left[\begin{array}{lcl} \text{pr.x} & : & \mathit{Ind}\\ \text{pr.c}_{\text{named}} & : & \text{named(pr.x, “Sam”)}\\ \text{pr.pr} & : & \mathit{Rec} \end{array}\right]$$

What we would like is for (12a) to be a subtype of (12b), that is, any situation of type (12a) is also of type (12b). But this is manifestly not the case. The labels do not match, for one thing. And yet intuitively there should be a match here. (12b) requires that there is an individual named Sam in any situation of the type and so does (12a). Our intuition rests on the equivalences of relabelling and flattening records that Cooper (2012) discusses.[6] We extend this flattening to types. If we flatten (12a), using complex labels so that we can get back to the unflattened type if we want, we obtain (12c). If we flatten and relabel (12b) with appropriate labels from (12c) we can obtain (12d). (12c) is a subtype of (12d).[7]

Let us summarize what we have done here a little more formally. We will use η as a variable over relabellings of a type and say that $\eta(T)$ is the result of relabelling T by η. We will use $\varphi(T)$ to represent the result of flattening T and φ^{-} to represent the inverse of flattening. (Thus $\varphi^{-}(\varphi(T)) = T$.) If *f* is a function returning types, a dependent type, we will use $\mathfrak{F}(f)$ to denote the fixed point type of *f* following (Cooper, 2012).[8] This is the type obtained by merging the domain type of the function with the type it returns, adjusting labels as necessary. If $f : T_1 \rightarrow T_2$ is an update function and T is a type (corresponding to commitments on the gameboard), then f can update T iff there is some relabelling η of $\varphi(T_1)$ such that $\varphi([\text{pr} : T]) \sqsubseteq \eta(\varphi(T_1))$. The result of updating T with f is then $\varphi^{-}(\varphi([\text{pr} : T]) \mathbin{\underset{.}{\wedge}} \eta(\varphi(\mathfrak{F}(f))))$. Suppose we want to update T with (9c). We first check that the flattening of [pr:T] is a subtype of some relabelling, η, of the flattening of (12b), that is the domain type of the function. If this holds, then we can update by merging the two flattened types and then reversing the flattening.

What happens if a match is not found and we are therefore unable to update the gameboard? Then accommodation must take place. We assume the kind of model discussed in Cooper and Larsson (2009) and Larsson and Cooper (2009) where not only a gameboard is present (a kind of short term memory) but also resources (a kind of long term memory). We think of one kind of accommodation as finding a match in the resources and “loading” this into the gameboard. If we think of resources as providing a record type or a collection of record types modelling long term memory, then the accommodation process could build on the techniques we have described here for update. The accommodation would involve first updating the gameboard with a subtype of the type required by the dialogue contribution we are trying to integrate. This subtype would be derived from the resources. It seems reasonable to suppose that the type found should be a proper subtype of the one required by the utterance, that is a type which provides more information that is presupposed by the utterance. This seems important to model intuitive notions of “identifying” objects,

---

[6] Flattening in TTR is conceptually related to the notion of path equation in Lexical Functional Grammar (Dalrymple et al., 1995).

[7] The order of the fields in our notation is not significant since record types are modelled as sets of ordered pairs (Cooper, 2012).

[8] Note that while there may not always be a fixed point for such a function, that is, some a such that $a : f(a)$, there will be a fixed point *type*, which may be empty.

that is, being able to provide further information about them. Now the gameboard will meet the requirements of the dialogue contribution we are trying to integrate and we can proceed with the update. Another kind of accommodation can be used in a situation where the resources do not provide an appropriate type. This involves updating the gameboard with the required type even though you did not have a match for it. Our suggestion would be that the update algorithm first looks for a match on the gameboard, then if that fails, in the resources and if that fails too, simply adding the required type. This process must also interact with clarification strategies. Clarification may be used either in the case where no match is found or more than one match is found. Of course, there can be other factors involved besides simply finding a match. For example, you may have reference to a person named Sam in your resources but you know that there is no way that your interlocutor could know about that particular Sam – thus it is a matter not just of finding a match but also an appropriate match or at least a match for which you do not have evidence that it is inappropriate.

3 Definite descriptions

For orientation, we will start our discussion of definite descriptions by mimicking Montague's (1973) treatment. We will use (13a) to represent the property of being a dog, that is, (13b).

(13) a. dog′

b. $\lambda r{:}\left[\text{x:}\mathit{Ind}\right].\ \left[\text{e:dog}(r.\text{x})\right]$

c. $\lambda P{:}\mathit{Ppty}.\ \left[\text{e:the}(\text{dog}',\text{P})\right]$

d. $[\downarrow P]$ =
$\{a \mid \exists r[r : \left[\text{x:}\mathit{Ind}\right] \wedge r.\text{x} = a \wedge [\check{\ }P(r)] \neq \emptyset]\}$
where for any type T,
$[\check{\ }T] = \{a \mid a : T\}$

e. $[\check{\ }\text{the}(P, Q)] \neq \emptyset$ iff
$\mid [\downarrow P] \mid = 1$ and $[\downarrow P] \subseteq [\downarrow Q]$

Then Montague's generalized quantifier treatment of definite descriptions is exemplified by (13c). This is the treatment of generalized quantifiers in TTR presented by Cooper (2011) and Cooper (2013). If P is a property, then we use $[\downarrow P]$ to represent the set of individuals that have P, as defined in (13d). Then we can say that the(P,Q) is a non-empty situation type (is "true") just in case $[\downarrow P]$ has exactly one member and $[\downarrow P]$ is a subset of $[\downarrow Q]$, as stated in (13e). This is a variant of the Russellian treatment of definite descriptions. It does not have any presuppositional element, that is, in our terms, it does not place any requirements on the interlocutor's gameboard in order to allow update. Furthermore it requires uniqueness apparently *tout court* rather than limited to a particular situation.

We fix the second problem first by introducing a resource situation (Barwise and Perry, 1983; Cooper, 1996). We allow properties to be restricted to a particular situation. Thus dog′$\upharpoonright s$ will be used to represent the property of being a dog in s as defined in (14a).

(14) a. $\lambda r{:}\left[\text{x:}\mathit{Ind}\right].\ \left[\text{e=}s\text{:dog}(r.\text{x})\right]$

b. $\lambda r{:}\begin{bmatrix}\text{s:}\mathit{Rec}\\ \text{e:unique}(\text{dog}',\text{s})\end{bmatrix}.$
$\lambda P{:}\mathit{Ppty}.\ \left[\text{e:every}(\text{dog}'\upharpoonright r.\text{s},P)\right]$

c. $[\check{\ }\text{unique}(P, s)] \neq \emptyset$ iff $\mid [\downarrow P \upharpoonright s] \mid = 1$

This notion of resource situation can then be exploited in an update interpretation for the definite description as in (14b). Here the predicate 'unique' is characterized as in (14c). That is, unique(P,s) holds just in case the set of individuals which have the property P restricted to s has exactly one element. This interpretation can combine with the interpretation of *left* and be matched against the commitments type on an agent's gameboard in an exactly similar fashion to that discussed in the preceding section.

We now consider the treatment of intensional verbs such as *look for*. If we follow Montague's original treatment the interpretation of *Sam is looking for the trainset* would involve the type of situations in (15a). If we adjust this using the kind of update functions we have suggested and the combination rule corresponding to functional application suggested in the previous section, we would obtain (15b). This is incorrect for the intensional reading since it requires the interlocutor to find a relevant situation with a unique trainset but in (3) both dialogue participants know that there is no trainset. We might then take inspiration from Montague's intensional analysis and say that the second argument to 'look_for' is the update function itself, not the result of applying it to a context. This would give us (15c). This seems hopeful, but it is still not quite right. Now *the trainset*

(15)

a. $$\left[\begin{array}{l}\text{e:look\_for(sam,}\\ \qquad \lambda P{:}\textit{Ppty}.\ \left[\text{e:the(trainset}',P)\right])\end{array}\right]$$

b. $$\lambda r{:}\left[\begin{array}{l} \text{f:}\left[\begin{array}{l}\text{x:}\textit{Ind}\\ \text{e:named(f.x, “Sam”)}\end{array}\right]\\ \text{a:}\left[\begin{array}{l}\text{f:}\textit{Rec}\\ \text{a:}\left[\begin{array}{l}\text{s:}\textit{Rec}\\ \text{e:unique(trainset}',\text{a.a.s)}\end{array}\right]\end{array}\right]\end{array}\right].$$
$$\left[\begin{array}{l}\text{e:look\_for(}r\text{.f.x, }\lambda P{:}\textit{Ppty}.\\ \qquad \left[\text{e:every(trainset}'\upharpoonright r\text{.a.a.s, }P)\right])\end{array}\right]$$

c. $$\lambda r_1{:}\left[\begin{array}{l} \text{f:}\left[\begin{array}{l}\text{x:}\textit{Ind}\\ \text{e:named(f.x, “Sam”)}\end{array}\right]\\ \text{a:}\left[\begin{array}{l}\text{f:}\textit{Rec}\\ \text{a:}\textit{Rec}\end{array}\right]\end{array}\right].$$
$$\left[\begin{array}{l}\text{e:look\_for(}r_1\text{.f.x,}\\ \qquad \lambda r_2{:}\left[\begin{array}{l}\text{s:}\textit{Rec}\\ \text{e:unique(trainset}',\text{a.a.s)}\end{array}\right].\\ \qquad\quad \lambda P{:}\textit{Ppty}.\\ \qquad\qquad \left[\text{e:every(trainset}'\upharpoonright r_2\text{.s, }P)\right])\end{array}\right]$$

d. $$\lambda r_1{:}\left[\begin{array}{l} \text{f:}\left[\begin{array}{l}\text{x:}\textit{Ind}\\ \text{e:named(f.x, “Sam”)}\end{array}\right]\\ \text{a:}\left[\begin{array}{l}\text{f:}\textit{Rec}\\ \text{a=}\left[\begin{array}{l}\text{s:}\textit{Rec}\\ \text{e:unique(trainset}',\text{s)}\end{array}\right]{:}\textit{RecType}\end{array}\right]\end{array}\right].$$
$$\left[\begin{array}{l}\text{e:look\_for(}r_1\text{.f.x,}\\ \qquad \lambda r_2{:}r_1\text{.a.a}\ .\\ \qquad\quad \lambda P{:}\textit{Ppty}.\\ \qquad\qquad \left[\text{e:every(trainset}'\upharpoonright r_2\text{.s, }P)\right])\end{array}\right]$$

e. if $\alpha : (T_1 \rightarrow ((T_4 \rightarrow T_2) \rightarrow T_3))$ and $\beta : (T_4 \rightarrow T_2)$ then the *combination of α and β based on intensional functional application* is

$$\lambda r_1{:}\left[\begin{array}{l}\text{f:}{T_1}^{\text{f.}}\\ \text{a=}T_4{:}\textit{Type}\end{array}\right].\ \alpha(r_1.\text{f})(\lambda r_2 : r_1.a\ .\ \beta(r_2))$$

equivalently:

$$\lambda r_1{:}\left[\begin{array}{l}\text{f:}{T_1}^{\text{f.}}\\ \text{a=}T_4{:}\textit{Type}\end{array}\right].\ \alpha(r_1.\text{f})(\beta)$$

is not placing any requirement on the interlocutor's gameboard and yet our intuitions and the evidence from (3) suggest that the interlocutor needs to know “which trainset is being talked about” without this entailing a commitment to there being such a trainset. How can this be? If we talk in presupposition terms ‘look_for’ is behaving as a hole according to (15b) and as a plug according to (15c). The remaining option is that it behaves as a filter, that is, it projects up a modification of the presupposition associated with *the trainset*. Our analysis will say that instead of projecting up the requirement that there is a situation with a unique trainset, the projected requirement is that a type of situation with a unique trainset is available on the gameboard without any requirement of there being something of this type. The analysis is given in (15d). The general combination rule, which should be compared with (9b), is given in (15e).[9]

In (3), dialogue participant B is clearly in accommodation mode, and, following our discussion at the end of the previous section, in difficulty trying to find a proper subtype of the type required by *the trainset* – hence, B's clarification request. Intuitively, A's response to the clarification request should provide that subtype, though at the time of writing it is a little unclear how we show technically that the interpretation of the noun-phrase provides a subtype. At this point in the dialogue, we suggest, the discussion is subordinated to that type which is to be placed on the gameboard. Thus it is that type which for the period of the subordination as if it is the type representing all the commitments on the gameboard. Once the type has been specified it will be inserted as a type in a field in the commitments type, requiring the type to be available but not requiring that there be anything of the type. This seems to provide a way of thinking about a number of different examples of intensional identity across dialogue turns, though working out the exact details of the mechanisms involved belongs to the realm of future work.

4 Conclusion

We have proposed a compositional treatment of proper names and definite descriptions using TTR

---

[9] In a fully explicit treatment this would require α to have a polymorphic type since in general the domain type of β could be any record type thus requiring α to apply to functions from a range of types. A similar situation in connection with the analysis of generalized quantifiers is discussed in Cooper (2011).

which makes a tight coupling between a compositional update semantics and the theory of dialogue gameboards. We have suggested that this provides a rather natural treatment of an otherwise puzzling phenomenon when definite descriptions are embedded below intensional verbs. We have sketched how this compositional semantics could interact with a theory of accommodation and clarification interactions, though this part of the theory is still in need of technical development.

Acknowledgments

I would like to thank Liz Coppock for useful discussion of this paper and three anonymous referees for insightful comments. This research was supported in part by VR project 2009-1569, Semantic analysis of interaction and coordination in dialogue (SAICD).

References

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. Bradford Books. MIT Press, Cambridge, Mass.

Robin Cooper and Staffan Larsson. 2009. Compositional and ontological semantics in learning from corrective feedback and explicit definition. In Jens Edlund, Joakim Gustafson, Anna Hjalmarsson, and Gabriel Skantze, editors, *Proceedings of DiaHolmia: 2009 Workshop on the Semantics and Pragmatics of Dialogue*, pages 59–66. Department of Speech, Music and Hearing, KTH.

Robin Cooper. 1996. The role of situations in generalized quantifiers. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*. Blackwell, Oxford.

Robin Cooper. 2011. Copredication, Quantfication and Frames. In Sylvain Pogodalla and Jean-Philippe Prost, editors, *Logical Aspects of Computational Linguistics: 6th International Conference, LACL 2011*, number 6736 in Lecture Notes in Artificial Intelligence, pages 64–79. Springer.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics, pages 271–323. Elsevier BV. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Robin Cooper. 2013. Clarification and Generalized Quantifiers. *Dialogue and Discourse*, 4(1):1–25.

Elizabeth Coppock and David Beaver. 2012. Weak uniqueness: The only difference between definites and indefinites. In *Proceedings of SALT*, volume 22.

Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell, and Annie Zaenen, editors. 1995. *Formal Issues in Lexical Functional Grammar*. Number 47 in CSLI Lecture Notes. CSLI Publications.

Paul Elbourne. 2012. *Definite Descriptions*. Clarendon Press, Oxford.

Chris Fox and Shalom Lappin. 2005. *Foundations of Intensional Semantics*. Blackwell Publishing.

Jonathan Ginzburg and Robin Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3):297–365.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In Harry Bunt, editor, *Proceedings of the 1st International Workshop on Computational Semantics*, Tilburg University. ITK Tilburg.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Alexander Grosu and Manfred Krifka. 2007. *The gifted mathematician that you claim to be*: Equational intensional 'reconstruction' relatives. *Linguistics and Philosophy*, 30:445–485.

Irene Heim. 1982. *The Semantics of Definite and Indefinite NPs*. Ph.D. thesis, University of Massachusetts at Amherst.

Sarah Hulsey and Uli Sauerland. 2006. Sorting Out Relative Clauses. *Natural Language Semantics*, 14:111–137.

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

Hans Kamp. 1981. A Theory of Truth and Discourse Representation. In Jeroen Groenendijk, Theo Janssen, and Martin Stokhof, editors, *Formal Methods in the Study of Language*, number 135 in Mathematical Centre Tracts. Mathematisch Centrum, Amsterdam.

Lauri Karttunen. 1973. Presuppositions of compound sentences. *Linguistic Inquiry*, 4(2):pp. 169–193.

Staffan Larsson and Robin Cooper. 2009. Towards a formal view of corrective feedback. In Afra Alishahi, Thierry Poibeau, and Aline Villavicencio, editors, *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 1–9. EACL.

Staffan Larsson and David R. Traum. 2001. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3&4):323–340.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, University of Gothenburg.

Richard Montague. 1973. The Proper Treatment of Quantification in Ordinary English. In Jaakko Hintikka, Julius Moravcsik, and Patrick Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 247–270. D. Reidel Publishing Company, Dordrecht.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven. ed. and with an introduction by Richmond H. Thomason.

Matt Purver and Jonathan Ginzburg. 2004. Clarifying noun phrase semantics. *Journal of Semantics*, 21(3):283–339.

Craige Roberts. 1987. *Modal Subordination, Anaphora, and Distributivity*. Ph.D. thesis, University of Massachusetts at Amherst.

Robert C. Stalnaker. 1978. Assertion. In Peter Cole, editor, *Pragmatics*, volume 9 of *Syntax and semantics*. Academic Press.

Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy*, 25:701–721.

Richmond Thomason. 1980. A model theory for propositional attitudes. *Linguistics and Philosophy*, 4:47–70.

Henk Zeevat. 1989. Realism and Definiteness. In Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types and Meanings*, volume II: Semantic Issues. Kluwer, Dordrecht.

Visual and linguistic predictors for the definiteness of referring expressions

Manjuan Duan, Micha Elsner and **Marie-Catherine de Marneffe**
Linguistics Department
The Ohio State University
Columbus, OH, 43210, USA
`{duan,melsner,mcdm}@ling.osu.edu`

Abstract

This study explores the impact of visual context on the conceptual salience of a discourse entity, using descriptions of how to find specific targets in cartoon scenes. Significant positive correlation is observed between larger and more salient objects and definite expressions, whereas more cluttered images are positively related to indefinite expressions. Incorporating these findings with other linguistic factors, we build a mixed-effects logistic regression model for predicting referring forms. The model reaches 62% accuracy. This study helps us to understand better how physical context, like an image, determines the linguistic properties of a discourse.

1 Introduction

When presented with a picture, how do you start your description? How will visual factors affect the expressions you use? And how do these factors interact with contextual and discourse features? Answering these questions will help to build a connection between the visual clues we perceive from a picture and the particular linguistic expressions we choose to describe it. It will also facilitate referring expression generation (REG) (Krahmer and van Deemter, 2012): the task of generating natural and contextually proper referring expressions.

In this study, we examine the roles played by visual features of an object and its visual context in determining whether, in a description, it will be mentioned by a simple definite NP, a long, descriptive definite expression, an indefinite, a demonstrative or a pronoun. We find that visual features like area and low-level visual salience are positively associated with definite referring expressions as a whole, suggesting that visually prominent objects are treated as more conceptually salient when we describe them.

These results are important for two reasons. First, they draw a firm connection between linguistic theories of reference which appeal to salience, and the low-level perceptual mechanisms from which salience arises. By doing so, they help to situate these theories in the wider context of cognitive science. Secondly, there is comparatively little research investigating the effect of visual properties on referring forms. Most previous research on text generation for REG has focused on content selection (Krahmer and van Deemter, 2012) where several studies have found effects of visual salience. These results suggest that human speakers also take vision into account during sentence planning and realization.

We use mixed-effects regression models to analyze the importance of several visual and linguistic factors in a corpus of visual-scene referring expressions (Clarke et al., 2013). These models can be used to predict referring forms on new data. Our classifier achieves 62% accuracy, which is 30% better than the majority baseline, and 6% better than a classifier without visual features, demonstrating its usability for generating contextually appropriate referring expressions for visual scenes.

2 Background

2.1 Linguistic theory

Linguists have proposed many different theories to account for the relationship between the salience of discourse entities and the kinds of referring expressions that can be used to describe them. Although different terms, such as topicality, givenness, accessibility, prominence, familiarity or salience, are used, they all converge on one point: referring expressions reflect the cognitive status of discourse entities they refer to (Prince, 1999; Chafe, 1976; Givon, 1983; Gundel et al., 1993; Ariel, 1988; Roberts, 2003).

Several of these theories match the cognitive or attentional states associated with a discourse entity and specific linguistic forms of reference. Gundel et al. (1993) use the term *givenness* to illustrate how salient a discourse entity is. They specify a scale of attentional states corresponding to different forms of referring expressions. The Givenness Hierarchy they suggest is: *in focus* > *activated* > *familiar* > *uniquely identifiable* > *referential* > *type identifiable*. Related to these six cognitive statuses are the forms of referring expressions that these statuses license: *it* > *that, this, this N* > *that N* > *the N*> *indefinite this N* > *a N*. Each status on the hierarchy is a necessary and sufficient condition for the appropriate use of a linguistic form. For example, a discourse entity has to be *in focus* to be referred by a pronoun, or a discourse entity has to be *uniquely identifiable* to license the definite expression *the N*.

Along the same line, Roberts (2003) proposes that the use of a definite NP presupposes that that the NP is *familiar* (i.e., that there is a corresponding discourse referent already in the discourse context), and that this discourse referent is *unique* among the discourse referents in the context. She also further differentiates *familiarity* into *strong familiarity* and *weak familiarity*. Strong familiarity is reserved for the more commonly assumed notion of familiarity, where it usually involves explicit previous mention of the entity in question, while an entity is weakly familiar when its existence is *entailed* by the local context. Hence, weak familiarity subsumes strong familiarity but is more inclusive, including discourse referents introduced non-linguistically, on the basis of contextual entailments (including perceptually accessed information) alone.

Ariel (1988; 1991) proposes a similar theory in which the complexity of referring forms reflects their accessible status in our mind. Basically, more reduced linguistic forms suggest more accessible or more salient status in the discourse. Based on her empirical study, she proposes a graded Accessibility Marking Scale in which she differentiates nominal descriptions with modifiers and those without modifiers. In general, expressions with modifiers refer to entities with lower accessibility. For example, short definite expressions denote discourse referents that are more accessible than long ones; the descriptive content of long definites helps to further single out their discourse referents.

These theories are attractive to us because they make efforts to capture the correlation between cognitive status on the one side and linguistic forms of referring expression on the other side.

What is generally missing is a fully grounded theory which explains how low-level percepts affect the cognitive status ranking. While it is universally acknowledged that non-linguistic factors play a role, most research has focused on linguistic features which can create or indicate a high cognitive status for an entity: for instance, Grosz et al. (1995) proposes a ranking scale of grammatical roles played by the discourse entities, *subject* > *object* > *others*, see also (Kameyama, 1986; Hudson et al., 1986). Other factors like the distance between the entity and its previous mention, the competition from other discourse entities and the (in)animacy of the discourse entities have also been studied as cues to determine the cognitive status of a discourse entity (Hobbs, 1976; Mitkov, 1998; Haghighi and Klein, 2010). When present, these linguistic features are highly influential, often overriding non-linguistic perceptual factors (Viethen et al., 2011a). But when they are not, less is known about which perceptual features matter in selecting appropriate referring forms.

2.2 Referring expression generation

Both psycholinguists and text generation specialists have examined precisely the case in which visual information has the greatest influence: one-shot referring tasks (i.e., without discourse context) involving an object in a visual scene.

Viethen et al. (2011b) analyze a corpus of map-task dialogues and find that visual context is *not* an important factor in deciding content of a referring expression, even for first mentions. However, other studies have found effects for visual features.

Kelleher et al. (2005) claim that salience—both visual and linguistic—is an important overarching semantic category structuring visually situated discourse. They describe a system which uses simple measurements of visual salience—bounding box area and distance to screen center—for both language understanding and REG content selection, and find these features are helpful. Duckham et al. (2010) use a variety of visual and perceptual features to select landmarks for computer-generated navigation instructions.

Clarke et al. (2013) also find a role for visual features in content selection. (They argue that the discrepancy with Viethen et al. might be accounted for by the stimuli—the images Clarke et al. use are more complex.) They find that visual properties (salience, clutter, area, and distance) influence referring expression generation for targets embedded in images from "Where's Wally?" books. Referring expressions for large target objects are shorter than those for small targets, and expressions about targets in highly cluttered scenes use more words. Also, people are more likely to choose large, salient objects which are close to the target as landmarks in relational descriptions.

Comparatively fewer studies have investigated how low-level visual features affect linguistic forms. Montag and MacDonald (2011) examined how visual salience affects the linguistic structure choice in terms of passive or active voice in relative clauses.

Closer to our work, Vogels et al. (2013) study how visual salience affects the choice of referent and the choice of referring forms when interacting with linguistic context in two story-completion experiments. They find that visual salience influences the choice of referent and does so independently of linguistic salience. But visual salience does not affect the choice of referring forms, which are strongly affected by linguistic salience. They conclude that visual salience has an influence on the global interpretation of the scene, but does not directly affect the accessibility status of individual entities— that is, people use different types of information in choosing a referent and choosing a referring expression.

In contrast, we do find effects from visual information on referring form, but nonetheless, we believe our study accords with Vogels et al. (2013). In their study, the two possible linguistic forms considered are pronouns and full noun phrases. Pronouns are a referring form which is highly sensitive to linguistic context, and our results also show they are relatively insensitive to visual effects; our strongest effects are in distinguishing different types of NP. Moreover, our one-shot referring task provides no linguistic context to begin with, while the story completion task of Vogels et al. (2013) provides previous referring expressions for the entities in all experimental conditions.

All the research introduced above shows that salient landmarks are more likely to be chosen in route description or scene descriptions than less salient ones and salient objects are more likely to be chosen as subject referent, which establishes the important role that visual salience plays in content selection. Both Montag and MacDonald (2011) and Vogels et al. (2013) study how visual salience affect our choice of concrete linguistics forms, but these studies involve highly controlled experimental environments in which perceptual variables are manipulated in a fairly coarse way, so that visual salience can be considered as a categorical variable rather than a continuum. Moreover, although Vogels et al. (2013) considers the choice of pronouns vs NPs , they leave open the issue of definiteness: what kind of NP to produce.

In this paper, we reanalyze Clarke et al. (2013)'s data, investigating which visual features of an object in an image or visual properties of the image as a whole affect people's choice of concrete linguistic referring forms. This study not only reveals the effects of various perceptual factors but also quantifies their relative importance. We show that both visual characteristics of the referent (visual salience and size) and a characteristic of the image as a whole (clutter) correlate with increased use of definite expressions. Furthermore, since visual factors have measurable effects on people's choice of referring forms, then consideration of these factors in referring expression generation tasks should be beneficial.

2.3 Visual salience

The visual salience (*Salience*) of an object (Toet, 2011) is a description of how much the object stands out from the background. Perceptual psychologists have developed models of visual salience, which typically aggregate low-level features such as color and contrast, and compare the features around each point to those in the image in general in order to predict how different the point will look from its surroundings. The size and central location of an object are also important (Tatler, 2007). Such models can predict fixations during scene viewing (Itti and Koch, 2000). Re-

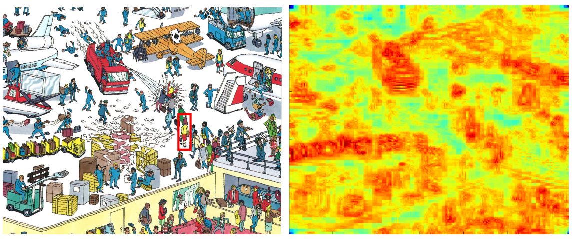


Figure 1: An image from our corpus and the corresponding visual salience map produced by the bottom-up component of Torralba et al. (2006); red indicates high salience scores, blue low salience scores.

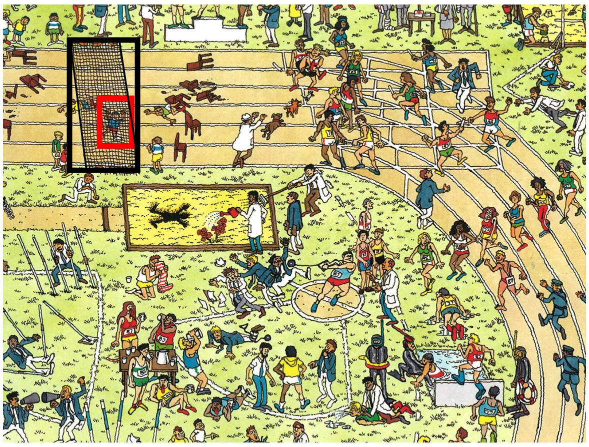
lated models from visual search (Wolfe, 1994) can also be used to predict how quickly subjects find a target object in a visual search task.

The Torralba et al. (2006) model used in our experiments is a typical contrast-based salience model (which we augment by including area, centrality and distance features as independent predictors).[1] It computes a visual salience score for each pixel in the image using a bank of oriented filters, then assigns a salience score to each bounding box which is the maximum over pixels it contains. The pixel scores are illustrated in Figure 1, which illustrates the visual prominence of the fire truck and the line of baggage handlers.

Visual clutter is a measurement of scene complexity; high clutter leads to difficulty when visually searching for objects (Henderson et al., 2009). Models of clutter (Rosenholtz et al., 2007) also depend on local image features such as color and orientation; in general, if these features are highly variable (many different colors and edge angles are represented), the scene will appear cluttered and hard to search.

3 Methods

We use a corpus collected in Clarke et al. (2013),[2] consisting of descriptions of specific target people in cartoon scenes from the children's book series "Where's Wally". The descriptions were elicited on Mechanical Turk, by asking participants to explain to someone else how to find a target person in the picture. Clarke et al. (2013) annotated the textual descriptions by marking references to visible objects and linked each one to a corresponding bounding box in the image. Their annotation scheme distinguishes two types of objects: the *target* is the person in the picture whom the subject was instructed to describe, while *landmarks* are other objects in the picture that the subject uses to describe the target. They also distinguish between textual mentions of landmarks that are part of a relative description ("near the bus") (Dale and Haddock, 1991), and those whose existence is *established* without giving a relative description ("look at the bus"). An example of the annotation is given in Figure 2.



`Under <lmark rel="targ" obj="imgID">a net</lmark> is <targ>a small child wearing a blue shirt and red shorts</targ>.`

Figure 2: An example image and RE from the corpus with the target marked by a red box. The annotator has added a black box for the landmark (in this case the net). Words describing the target and landmark in the RE are XML-tagged.

Our goal here is to characterize how visual features affect the way people perceive definiteness of a discourse entity and choose referring forms accordingly from a cognitive/linguistic standpoint. We therefore used the totality of the descriptions in the corpus, without conducting experiments to determine whether they would lead to a successful/quick identification of the target by the listener. The fact that we did not filter out such "bad/unsuccessful" descriptions might be a weakness as far as applications are concerned, but from the cognitive/linguistic investigation that concerns us, these descriptions are a valuable source of information about how speakers compose descriptions.

---

[1] The Torralba et al. (2006) model also includes a top-down component which models task-based attentional effects, but this is not used.

[2] `http://datashare.is.ed.ac.uk/handle/10283/336`

	Pron	**Demo**	**SDef**	**LDef**	**Indef**
Counts	575	213	1013	1584	1594
%	11.5	4.3	20.3	31.8	32.0

Table 1: Distribution of referring forms.

We distinguish six classes of referring form: pronouns, demonstratives, short definite NPs, long definite NPs, indefinite NPs and bare singulars. We manually annotate each tagged mention of a visual object with its appropriate class.[3] Demonstratives are NPs headed by *this, that, these* and *those*. Definite NPs are those headed by *the*. Short definite NPs are definite NPs without any modifiers and long definite NPs are those with modifiers like adjectives, prepositional phrases, and relative clauses. We split the definites in this way in order to investigate the Accessibility Marking theory of Ariel (1988). Indefinite NPs are those headed by *a, an, some* or plural nouns. Bare singulars are singular nouns not headed by any determiners, like "man with a hat" or "brown dog"; these are ungrammatical in standard English, but occur in Mechanical Turk elicitations. The corpus contains 447 bare singulars; a preliminary analysis using the features below showed that these were similar in their distribution to definites and usually misclassified as such. We conclude that the bare singular form is an alternate form of the definite, and in the rest of our analysis one-word bare singulars are merged with short definite NPs and longer bare singulars with long definite NPs (Table 1).

We perform one-vs-all mixed-effects logistic regression analyses with R (Bates et al., 2011). We incorporate random intercepts for speaker (N=115) and image (N=11), and three types of fixed-effects features: task-based, visual and linguistic.

Task-based features

The task features indicate whether the object being referred to is the target of the description (*Target*) or a landmark (*Lmark*).

Visual features

Visual features of the described object include its area (*Area*) as well as its centroid-to-centroid distance from the target (*Distance*). Another feature captures whether its bounding box overlaps with that of the target or, if it is a landmark in a relative description of some other object, with that object (*Overlap*) (Kelleher et al., 2005; Golland et al., 2010).

We also use two models from the perception literature as features in our analysis. Both of them are previously-implemented models from the perceptual psychology literature. We use the values computed and distributed by Clarke et al. (2013), which measure the visual salience of bounding boxes by using the bottom-up component of Torralba et al. (2006). We also compute visual clutter using two models proposed in Rosenholtz et al. (2007).[4] Feature congestion (*Congestion*) measures the variance in features like different colors, orientations, or luminance contrast changes in a given local area. Sub-band entropy (*Clutter* or *Clt*) measure represents the intuition that an "organized" scene is less cluttered. With more organization, and thus more redundancy, the brain (or computer) can represent an image with more efficient encoding, thus a lower value in this measure. It is inversely related to how many bits could be saved by JPEG-compressing the image (Rosenholtz et al., 2007; Asher et al., 2013). All the values of visual features used in this paper are distributed as part of the corpus.

Linguistic features

We use linguistic features found to be useful in previous studies of definiteness and information status (Nissim, 2006). In some cases we modified these feature definitions to rely on surface ordering rather than syntactic annotations, due to our lack of a parser for the Mechanical-Turk-elicited text.

Coref: We check if the phrase refers to a previously-mentioned entity, treating two phrases as coreferent if they resolve to the same bounding box in the image.

Establish: This feature captures whether the annotator marked the expression as *establishing* existence rather than part of a relative description, such as "look at the X", rather than a relative description like "near the X".

There-be: We have an explicit feature to capture *there+be* existential construction, known to disfa-

---

[3] The corpus also contains tags for non-visual objects ("the bottom left") and tags that are not mentions ("first on the left [implied *of X*]"); we exclude these from our analysis.

[4] We compute these scores ourselves, using the Matlab tools distributed by Rosenholtz.

Features	Pron	Demo	SDef	LDef	(Def)	Indef
Task						
Target	1.44 **	3.46 ***	0.60 *	-0.58 **	-0.003	-1.16 ***
Lmark	-0.74 ·	1.78 ***	1.07 ***	-0.86 ***	-0.09	0.22
Linguistic						
Coref	4.49 ***	0.75 ***	0.04	-1.61 ***	-0.09	-2.35 ***
There-be	-15.25	-15.43	-3.75 ***	-3.84 ***	-4.61 ***	5.33 ***
Be	-3.33 ***	-3.01 **	-2.11 ***	-2.77 ***	-2.99 ***	3.88 ***
First	0.89 ***	0.14	-0.50 **	-0.31 **	-0.54 ***	-0.41 **
Prep	-0.13	0.01	0.01	0.16 **	0.28 ***	-0.38 ***
Establish	0.55 *	2.16 ***	-0.17	-0.49 **	-0.73 ***	0.50 **
Visual						
Area	-0.35 ·	-0.81 *	0.64 ***	-0.38 ***	0.63 ***	-0.67 ***
Salience	-0.26 **	-0.01	0.08	0.05	0.11 *	-0.02
Overlap	0.001	0.47 ·	0.07	-0.4 ***	-0.46 ***	0.61 ***
Distance	0.16	-0.11	0.15 **	0.19 *	0.37 ***	-0.66 ***
Clutter	0.54	-0.17	0.01	-0.43 *	-0.37 *	0.34 *
Congestion	0.02	-0.21 ·	0.001	0.07	0.07	0.01
Interaction						
Target:Clt	-0.59	0.19	0.04	0.36 **	0.42 **	-0.37 *
Area:Clt	0.09	-0.54 *	-0.01	-0.09 *	0.39 ***	-0.47 ***
Salience:Clt	0.05	-0.05	0.28 ***	-0.07·	-0.11 *	0.15 **

Table 2: Coefficients learned by the one-vs-all mixed-effects models for predicting referring forms. Significance codes: p-value < 0.001, $* * *$; p-value < 0.01, $**$; p-value < 0.05, $*$; p-value < 0.1, $\cdot$. The model includes all pairwise interactions, but only significant interactions are shown. The "Def" column shows coefficients for a merged class containing both long and short definites.

vor definites (Ward and Birner, 1995).

Syntactic position: We checked whether the target is directly preceded by any form of *to be* (*Be*); whether it is directly preceded by a preposition (*Prep*) or whether it appears sentence-initially, a proxy for the subject grammatical role (*First*).

4 Results and analysis

The coefficients from our one-vs-all mixed effects logistic regression analysis are shown in Table 2.[5] The linguistic features generally behave as the existing literature leads us to expect. A previous coreferent mention has the expected impact on referring forms (Roberts, 2003): pronouns and demonstratives are favored as indicated by the positive estimate for *Coref*, whereas indefinites are disfavored (negative coefficient). Indefinite NPs are positively associated with *There* and *Be*. Definite NPs are positively related to *Prep*, indicating that uniquely identifiable discourse entities are more likely to be the complements of prepositions. *First* is positively related to pronouns, which supports the hypothesis that back-looking centers like pronouns tend to appear at linguistically salient positions like subject position to achieve better discourse coherence (Grosz and Sidner, 1986).

As for visual features, we find main effects of *Area* in favor of short definite NPs, against long definites and strongly against indefinites. This result accords with the Accessibility Marking Scale proposed by Ariel (1988), which uses short definites for more accessible objects, then long definites and finally indefinites.

The results for *Salience* are smaller, but appear to be similar. Visual salience has non-significant positive associations with both short and long definites; if both classes of definite are analyzed together, the effect reaches significance. We suspect the failure to find it with either subgroup is due to reduced power because of the relatively smaller datasets. Overall the results confirm our hypothesis that larger and more visually salient objects are also perceived as more prominent and tend to be referred to by definite expressions, especially short definites.

Overlap is positively related to indefinite expressions and *Distance* is positively related to definite expressions. A closer look will show that these two measures are inversely related; usually, when two objects are overlapped, the centroid distance between them is short. In other words,

---

[5] We also considered the distance of an object to the center of the image, but its effect was not significant.

Features	**Accuracy**	**Sig vs.**
Baseline (majority)	32.01	
Task features	38.92 ***	baseline
Linguistic features	54.68 ***	baseline
Visual features	42.19 ***	baseline
Task + visual features	43.30 ***	task
Task + ling features	56.11$^{\cdot}$	ling
Ling + visual features	58.08 ***	ling
Task + ling + visual	62.06 ***	ling + visual

Table 3: Prediction results for the different feature types, with distinction between short and long definite referring expressions. The last column indicates whether results significantly differ (Mann-Whitney U test).

Features	**Accuracy**	**Sig vs.**
Baseline (majority)	51.7	
Task features	55.32 ***	baseline
Linguistic features	72.44 ***	baseline
Visual features	55.94 ***	baseline
Task + visual features	56.78 ***	task
Task + ling features	73.27	ling
Ling + visual features	74.15	ling
Task + ling + visual	74.74	ling

Table 4: Prediction results for the different feature types with short and long definite expressions combined. The last column indicates whether results significantly differ (Mann-Whitney U test).

Gold ↓	Proposed →				
	Pron	Demo	SDef	LDef	Indef
Pron	454	17	32	62	10
Demo	49	26	28	108	2
SDef	44	11	398	463	97
LDef	63	3	157	1180	181
Indef	31	0	31	488	1044

Table 5: Confusion matrix for predicted referring forms.

speakers use more definite expressions to refer to objects far from the target of the description, while using more indefinites to refer to objects close by. Landmarks that are close by can be helpful even if they are hard to see (by helping the listener confirm that they have found the target). But distant landmarks must be easy to find in their own right, and this makes them better candidates for definite mentions.

Converging with the findings discussed above, the estimates for *Clutter* suggest that indefinite expressions are more likely to be used in a more crowded image. *Area* also interacts with *Clutter*: large objects are more likely to be definite and less likely to be indefinite when the image is more cluttered overall. This supports the results from linguistic research that indefinites need to be *type identifiable* (Gundel et al., 1993) while definites need to have uniquely identifiable referents (Gundel et al., 1993; Roberts, 2003). In an image where a lot of similar objects crowd together, many objects, especially smaller ones, will be hard to uniquely identify, so speakers may avoid using definite references for them. Alternatively, speakers might not be able to easily verify that the object is in fact unique in the image.

Using the predictions obtained from the five one-vs-all logistic regressions, we classify 479 randomly chosen NPs held out as test data, using the standard highest score strategy. Table 3 shows the classification accuracies. We find that all three types of features are significantly more effective than a majority baseline (always "indefinites"). Linguistic features are very robust in predicting referring forms as widely recognized by prior research, which itself improve the overall accuracy from 32% to 55%. Adding visual features also leads to significant improvement of predicting results on the top of baseline, linguistic features and task-based features, which gives stronger support for our hypothesis that low-level visual features play an important role in predicting linguistic forms for referring expressions. Our strongest model, using all feature sets together, scores 62%.

Table 4 shows the classification accuracies when short and long definite expressions are combined. All three types of features are still significantly more effective than the baseline majority, now definites. However, adding visual features does not lead to a significant improvement on top of linguistic and task-based features. This means combining short and long definite expressions reduces the prediction of visual features, which suggest visual features are most effective in differentiating short and long definite expressions.

Table 5 shows per-category prediction results for each of the referring forms, cross-validated over the entire dataset. Most pronouns are predicted to be pronouns, despite their low percentage in our data (11%); 16% are labeled as definites, and less than 2% as indefinites. Very few demonstratives (12%) are correctly predicted, since they are extremely under-represented in our data (4%). However, most of them are predicted as definites (64%) and pronouns (23%). 11% of definites are labeled as indefinite, showing that pronouns, demonstratives and definite expressions, as a group, share some common features, and our model draws a relatively sharp distinction between this group and the indefinites.

Although different cognitive states are proposed in linguistic research as necessary conditions for definite expressions, such as *uniquely identifiable* by Gundel et al. (1993) and *weak familiar* by Roberts (2003), all these theories claim that discourse entities which have higher cognitive status in the givenness scale, like *in focus* or *activated* can be referred to by either definite NPs or pro-forms like pronouns or demonstratives. We observe this predicted overlap in the usage of these three referring forms in the confusion results.

Of the remaining errors, we believe many are due to individual differences between speakers in terms of visual perception or describing style. Inspection of the random intercepts reveals that speakers vary in the overall proportions of different referring forms they use. In some cases this seems to be a matter of style: some people phrase their referring expressions as instructions ("Look for the man standing aside the red truck"), others describe ("A man standing...") and some use a telegraphic style ("man, in blue jeans, standing...").

Figure 3 also suggests that visual properties like "area" have different effects on people's choice of whether to use definite or indefinite expressions. Most subjects (lines curving sharply to the upper left) follow the general trend of using definite expressions for larger objects, but a few show weaker trends, or no trend at all. Whether the variance is caused by speakers perceiving the image differently, or reacting differently to visual factors, deserves future study.

5 Conclusion

In this study, we have revealed the correlation between the visual features of discourse entities and their referring forms. We find visual features like *area* and *salience* are positively related to definite expressions and indefinite expressions are more likely to be used in crowded images. Based on these findings, we train a classifier to predict the referring forms for these visual objects. Our classifier achieves 62% overall accuracy, 30% higher than the majority baseline. This study helps us to better grasp the interaction between linguistic properties of the discourse and the physical context in which utterances are grounded. In future work, we hope to incorporate these features into a full-scale surface realization system.

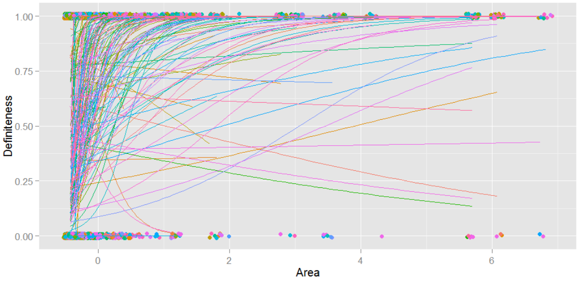


Figure 3: Logistic regression lines for proportion of definiteness as predicted by area for each of the 151 speakers in our data (data items shown as colored points). In general, larger area leads to more definite descriptions, but the effect varies across speakers and describing tasks.

Acknowledgements

We thank Jefferson Barlew, Alasdair Clarke, Gregory Kierstead, Hannah Rohde, Craige Roberts and the reviewers for their useful comments on earlier drafts of this paper.

References

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16:443–464.

Matthew F. Asher, David J. Tolhurst, Tom Troscianko, and Iain D. Gilchrist. 2013. Regional effects of clutter on human target detection performance. *Journal of Vision*, 13(5):1–15.

Doug Bates, Martin Maechler, and Ben Bolker. 2011. lme4: Linear mixed-effects models using S4 classes. Comprehensive R Archive Network.

Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C. Li, editor, *Subject and Topic*. Academic Press, New York.

Alasdair D. F. Clarke, Micha Elsner, and Hannah Rohde. 2013. Where's Wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology (Perception Science)*, Issue on Scene Understanding: Behavioral and computational perspectives.

Robert Dale and Nicholas J. Haddock. 1991. Generating referring expressions involving relations. In *EACL*, pages 161–166.

Matt Duckham, Stephan Winter, and Michelle Robinson. 2010. Including landmarks in routing instructions. *Journal of Location Based Services*, 4(1):28–52.

Talmy Givon. 1983. Topic continuity in discourse: an introduction. In T. Givon, editor, *Topic continuity in discourse: a quantitative cross-language study*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA, October. Association for Computational Linguistics.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, June.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California, June. Association for Computational Linguistics.

John M. Henderson, Myriam Chanceaux, and Tim J. Smith. 2009. The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of vision*, 9(1):1–8.

Jerry R. Hobbs. 1976. Pronoun resolution. Technical Report 76-1, City College New York.

Susan B. Hudson, Michael K. Tanenhaus, and Gary S. Dell. 1986. The effects of the discourse center on the local coherence of a discourse. In *Program of the Eighth Annual Conference of the Cognitive Science Society*.

Laurent Itti and Christof Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506.

Megumi Kameyama. 1986. A property-sharing constraint in centering. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*.

J. Kelleher, F. Costello, and J. van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167:62–102. Connecting Language to the World.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, March.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 869–875, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Jessica Montag and Maryellen MacDonald. 2011. How visual salience affects structure choice: Implications for audience design. In *Poster presented at the 24th Annual CUNY Conference on Human Sentence Processing*, Stanford, CA.

Malvina Nissim. 2006. Learning information status of discourse entities. In *Proceedings of EMNLP*, pages 94–102, Morristown, NJ, USA. Association for Computational Linguistics.

Ellen Prince. 1999. How not to mark topics: 'topicalization' in English and Yiddish. In *Texas Linguistics Forum*. University of Texas, Austin.

Craige Roberts. 2003. Uniqueness in definite noun phrases. *Language and Philosophy*, 26:287–350.

Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of Vision*, 7:1–21.

Ben W. Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17.

Alexander Toet. 2011. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2131 – 2146.

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. 2006. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 113:766–786.

Jette Viethen, Robert Dale, and Markus Guhe. 2011a. Generating subsequent reference in shared visual scenes: Computation vs re-use. In *Proceedings of* *the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1158–1167, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Jette Viethen, Robert Dale, and Markus Guhe. 2011b. The impact of visual context on the content of referring expressions. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 44–52, Nancy, France. Association for Computational Linguistics.

Jorrig Vogels, Emiel Krahmer, and Alfons Maes. 2013. Who is where referred to how, and why? the influence of visual saliency on referent accessibility in spoken language production. *Language and Cognitive Processes*, 28(9):1323–1349.

Gregory Ward and Betty Birner. 1995. Definiteness and the English existential. *Language*, 71(4):722–742, December.

Jeremy M. Wolfe. 1994. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1:202–238.

People hesitate more, talk less to virtual interviewers than to human interviewers

Lauren Faust[*] **and Ron Artstein**
USC Institute for Creative Technologies
12015 Waterfront Drive
Playa Vista CA 90094-2536, USA
`Lauren.Faust@student.csulb.edu` , `artstein@ict.usc.edu`

Abstract

In a series of screening interviews for psychological distress, conducted separately by a human interviewer and by an animated virtual character controlled by a human, participants talked substantially less and produced twice as many filled pauses when talking to the virtual character. This contrasts with earlier findings, where people were *less* disfluent when talking to a computer dialogue system. The results suggest that the characteristics of computer-directed speech vary depending on the type of dialogue system used.

1 Introduction

As computer dialogue systems become more commonplace, it becomes more relevant to ask how people's interaction with dialogue systems differs from interaction with other people. The answer, of course, will vary with different dialogue systems. This paper presents a study of comparable interview dialogues, where the interviewer is either a real person or an animated computer character controlled by a person ("Wizard of Oz" setting). Unlike earlier studies which showed that people hesitate less when talking to a computer, the present study shows that people hesitate more and produce twice as many filled pauses when talking to an animated conversational interviewer.

Existing studies show that qualitatively, human-computer dialogue exhibits many similarities to human-human dialogue. For example, an analysis of interactions between visitors and Max, an animated computer agent responding to typed input at the Heinz Nixdorf MuseumsForum in Paderborn, Germany in 2004, showed that visitors treated Max conversationally as a person, evidenced by conventional strategies of beginning and ending conversations and general cooperativeness (Kopp et al., 2005). Children have been shown to exhibit turn-taking behavior when interacting with a virtual peer (Sam the CastleMate: Cassell, 2004), and match their conversational style to that of a virtual character (Cassell et al., 2009). In an extensive literature review, Branigan et al. (2010) show that people align their speech patterns with computers at multiple levels of linguistic structure; this work also shows that the extent of alignment varies depending on whether the speaker thinks they are talking to a computer or to a person (though in the experiments cited, people were talking to computers in both belief conditions).

However, there are not many quantitative studies about the differences between comparable human-human and human-computer dialogues. Several early studies measured disfluencies in computer-directed speech. Oviatt (1995) looked at disfluencies in three corpora – a corpus of simulated human-computer interactions using speech and writing to accomplish transactional tasks such as paying bills or booking a rental car, a corpus of task-oriented telephone conversations regarding conference registration and travel arrangements, and a corpus of face-to-face dialogues and monologues giving instructions on how to assemble a water pump. The disfluency rate was significantly higher when talking to a person than when talking to a computer; within the computer-directed speech, disfluencies occurred at a higher frequency when the tasks were unconstrained rather than structured. Oviatt (1996) found that in a multimodal (speech + pen) map interaction task, disfluency rates were similar to those found in the computer-oriented speech from the previous study. Shriberg (1996) compared the frequency of disfluencies in three different corpora – a corpus of simulated human-computer interactions of air-travel planning, a corpus of real dialogues between travelers and travel agents, and the Switch-

---

[*] Now at California State University Long Beach

Interviewer Let's see, is there someone in your life that's been a really positive influence for you?
Participant Uh yeah, my husband, yeah.
Interviewer Yeah.
Interviewer What kind of values did you take away from him?
Participant Uh he's always uh thinking ahead and looks at the big picture and doesn't uh mull over trivial things so that's something that helped me.
Interviewer Mhm yeah, those are good traits to have.
Participant Yeah, yes.
Interviewer Um how did you guys meet?

Ellie Who's someone that's been a positive influence in your life?
Participant Uh my father.
Ellie Can you tell me about that?
Participant Yeah, he is a uh
Participant He's a very he's a man of few words
Participant And uh he's very calm
Participant Slow to anger
Participant And um very warm very loving man
Participant Responsible
Participant And uh he's a gentleman has a great sense of style and he's a great cook.
Ellie Uh huh
Ellie What are you most proud of in your life?

Figure 1: Example dialogues: face-to-face (left) and Wizard-of-Oz (right)

board corpus of general-domain telephone conversations. Here too, individuals were significantly more disfluent when talking to a person than when talking to a computer, producing more repetitions, deletions, and filled pauses. All the above studies found that the rate of disfluencies increased as the utterance length increased.

Though the aforementioned studies examine interactions between humans and computers, these dialogues cannot be said to mirror a face-to-face conversation. The computer partners are disembodied, communicating only with voice or with voice augmented by a graphical interface, and the dialogues are task-oriented rather than conversational. It is therefore not possible to draw inferences from these studies on how people will talk to conversational, embodied computer dialogue systems. Later studies used systems with more conversational characteristics: In Oviatt (2000), children aged 6–10 asked questions of computer images of sea animals, with rudimentary animations (blinking eyes) and synthesized speech; here too, the children were less disfluent when talking to the computer characters than when playing a 20-question game with an adult. A more realistic conversational agent was used in Black et al. (2009), where children aged 4–7 talked to an animated agent which used a combination of recorded and synthesized speech (Yildirim and Narayanan, 2009). In this study, children talking to the character exhibited disfluencies in fewer turns than when talking to an adult, though the effect was smaller than in the previous studies cited.

Other than Black et al. (2009) we have not found studies of comparable corpora of human-human and human-computer interaction with embodied conversational agents. The absence of such corpora is somewhat surprising, given that it has been known for several decades that people talk differently to computers and humans (e.g. Jönsson and Dahlbäck, 1988), and since human role-playing is often a preliminary step in developing conversational dialogue systems (e.g. Traum et al., 2008, section 4.3). The present study looks at a comparable corpus developed for such a purpose – a set of human-human interviews and character-human interviews in a Wizard-of-Oz setup, both collected for the eventual development of a fully automated conversational agent that will act as an interviewer, screening people for mental distress (see examples in Figure 1). In this corpus it turns out that the rate of filled pauses is higher when talking to a character than when talking to a person, suggesting that the previous results are not a general property of computer-directed speech, but rather specific to the type of dialogue systems used in the studies. The increase in disfluency when interviewed by an embodied conversational agent, compared to prior research showing a decrease in disfluencies when talking to disembodied agents, is consistent with the results of Sproull et al. (1996), who show that participants take longer to respond and type fewer words when interviewed by a talking face compared to a textual interview.

The remainder of the paper describes the corpus, the measures taken, and the differences found between human-human and character-human conversations. Our results show that patterns of conversation with disembodied, task-oriented dialogue systems do not carry over to embodied conversational agents. More generally, it is not appropriate to talk about how people talk to computers in general, because the way people talk varies with the type of dialogue system they talk to.

2 Method

2.1 Materials

We used a corpus of interviews, designed to simulate screening interviews for psychological distress, collected as part of an effort to create a virtual interviewer character. The interviews are of two types (see examples in Figure 1).

Face-to-face interviews, where a participant talks to a human interviewer situated in the same room; these are a subset of the interviews analyzed by Scherer et al. (2013) for nonverbal indicators of psychological distress.

Wizard-of-Oz interviews, where a participant talks to an animated virtual interviewer controlled by two human operators sitting in an adjacent room; a subset of these interviews were analyzed in DeVault et al. (2013) for verbal indicators of psychological distress.

The face-to-face interviews were collected during the summer of 2012. Participants were interviewed at two sites: at the USC Institute for Creative Technologies in Los Angeles, California, and at a US Vets site in the Los Angeles area. Participants interviewed at ICT were recruited through online ads posted on Craigslist.org; those interviewed at the US Vets site were recruited on-site, and were mostly veterans of the United States armed forces. After completing a set of questionnaires alone on a computer, participants sat in front of the interviewer for the duration of the interview (Figure 2); only the participant and interviewer were in the room. Interviews were semi-structured, starting with neutral questions designed to build rapport and make the participant comfortable, progressing to more specific questions about symptoms and events related to depression and PTSD (Post-Traumatic Stress Disorder), and ending with neutral questions intended to reduce any distress incurred during the interview. Participant and interviewer were recorded with separate video cameras, depth sensors (Microsoft Kinect), and lapel microphones. For additional details on the collection procedure, see Scherer et al. (2013).

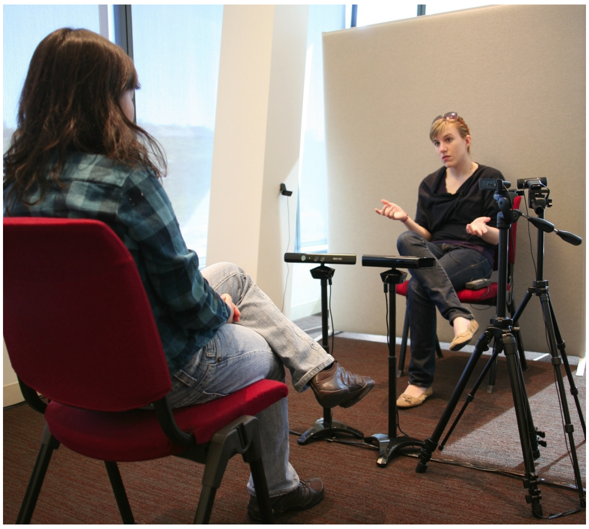


Figure 2: Face-to-face interview setup.

The Wizard-of-Oz interviews were collected in three rounds during the fall and winter of 2012–2013. All the participants were recruited through online ads posted on Craigslist and interviewed at the USC Institute for Creative Technologies. As with the face-to-face interviews, participants first completed a set of questionnaires on a computer, and then sat in front of a computer screen for an interview with the animated character, Ellie (Figure 3). No person other than the participant was in the room. The interviewer's behavior was controlled by two wizards, one responsible for the non-verbal behaviors such as head-nods and smiles, and the other responsible for verbal utterances (the two wizards were the same people who served as interviewers in the face-to-face data collection). The character had a fixed set of verbal utterances, pre-recorded by an amateur actress (the wizard controlling verbal behavior). The Wizard-of-Oz interviews were semi-structured, following a progression similar to the face-to-face interviews. Participants were recorded with a video camera, Microsoft Kinect, and a high-quality noise-canceling headset micro-

Figure 3: Ellie, the virtual interviewer.

phone.

There were small differences in protocol between the three rounds of the Wizard-of-Oz data collection. In the first round, the introductory explanation given to the participants did not explicitly clarify whether the interviewer character was automated or controlled by a person; in the subsequent rounds, each participant was randomly assigned to one of two framing conditions, presenting the character as either an autonomous computer system or a system controlled by a person. We did not find differences between the framing conditions on the measures described below, so the results reported in this paper do not look at the framing condition variable. An additional difference between the three Wizard-of-Oz collection rounds was the interview protocol, which became stricter and more structured with each successive round. Finally, with each round the character received a few additional utterances and nonverbal behaviors.

In both the face-to-face and Wizard-of-Oz conditions, each participant completed a series of questionnaires prior to the interview; these included the PTSD Checklist – Civilian Version (PCL-C) (Blanchard et al., 1996) and the Patient Health Questionnaire, depression module (PHQ-9) (Kroenke and Spitzer, 2002). There are strong correlations between the results of the two questionnaires (Scherer et al., 2013, Figure 1), so for the purpose of the analysis in this paper, we collapse these into a single assessment of distress: participants who scored positive on either of the questionnaires are considered distressed, while those who scored negative on both are considered non-distressed. In the face-to-face condition, interviewers received the results of the questionnaires prior to the interview, whereas in the Wizard-of-Oz condition, wizards were blind to the participant's distress condition.

Condition	Distressed	Non-distressed
Face-to-face	34	40
Wizard-of-Oz	59	124

Table 1: Participants and conditions.

Overall, our analysis considers the gross division of the participant population into two interview conditions (face-to-face and Wizard-of-Oz) and two distress conditions (distressed and non-distressed); see Table 1. We do not consider differences between the Wizard-of-Oz collection rounds or framing conditions, nor differences between the veteran and non-veteran populations or the individual interviewers in the face-to-face interviews. While it is known that demographic factors affect language behavior, and in particular disfluency rates (Bortfeld et al., 2001), differences between the US Vets and general population turned out non-significant on all the measures reported below, with the exception of rate of plural pronouns which was marginally significant at $p = 0.03$. Splitting the participant population into smaller groups would make it more difficult to detect the trends in the broad categories.

All the dialogues were segmented and transcribed using the ELAN tool from the Max Planck Institute for Psycholinguistics (Brugman and Russel, 2004),[1] and each transcription was reviewed for accuracy by a senior transcriber. Utterances were defined as continuous speech segments surrounded by at least 300 milliseconds of silence. For the face-to-face dialogues, both participant and interviewer were transcribed; for the Wizard-of-Oz dialogues only the participant was transcribed manually, while the interviewer utterances were recovered from the system logs.

2.2 Procedure

Several measures were extracted from the transcriptions of the interviews using custom Perl scripts.

Quantity measures: Total time of participant

---

[1] http://tla.mpi.nl/tools/tla-tools/elan

speech; total number of participant words; speaking rate; utterance length.

Disfluency measures: Filled pauses (*uh, um, mm*) per thousand words; percentage of utterances beginning with a filled pause.

Lexical items: First person singular (*I, me, our*) and plural (*we, us, our*) pronouns; definite (*the*) and indefinite (*a, an*) articles.

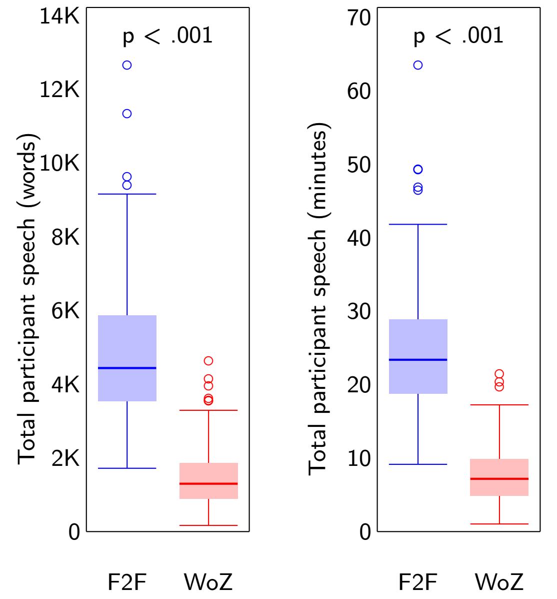
The above measures were calculated individually for each participant; we then compared the measures according to the 2×2 setup (interview condition and distress condition) described above for Table 1. Most of the significant effects we found are main effects of interview condition. Since the values typically do not follow a normal distribution, we report these effects using Wilcoxon rank-sum tests.

3 Results

3.1 Speech quantity

The face-to-face dialogues were substantially longer than Wizard-of-Oz dialogues (Figure 4): the median face-to-face dialogue participant uttered 4432 words and spoke for 23 minutes, while the median Wizard-of-Oz dialogue participant uttered only 1297 words and spoke for only 7 minutes; the differences are highly significant ($W \approx 450$, $n_1 = 74$, $n_2 = 183$, $p < .001$). The difference in speech quantity is likely due to several limitations of the wizard system. With a fixed set of utterances, the wizard runs out of things to say at some point, whereas human interviewers can engage the participants for much longer. Additionally, the human interviewer can tailor the questions to the participant's previous response, going deeper into each discussion topic than is possible for a wizard.

Not only did participants talk more in the face-to-face condition, they also used longer utterances. We calculated the mean number of words per utterance for each speaker (Figure 5, left panel): the median is 16 words per utterance in the face-to-face dialogues and 8 in the Wizard-of-Oz dialogues ($W = 1398$, $p < .001$). One possible reason for the difference is that speakers may be aligning their utterances to match the length of the interviewer's utterance. Another possible reason is the near-absence of verbal backchannels



Boxplot legend

Heavy line	Median
Shaded area	First and third quartiles
Whiskers	Data within 1.5 times interquartile range
Circles	Outliers beyond 1.5 times IQR
P-value	Wilcoxon rank-sum test

Figure 4: Speech quantity.

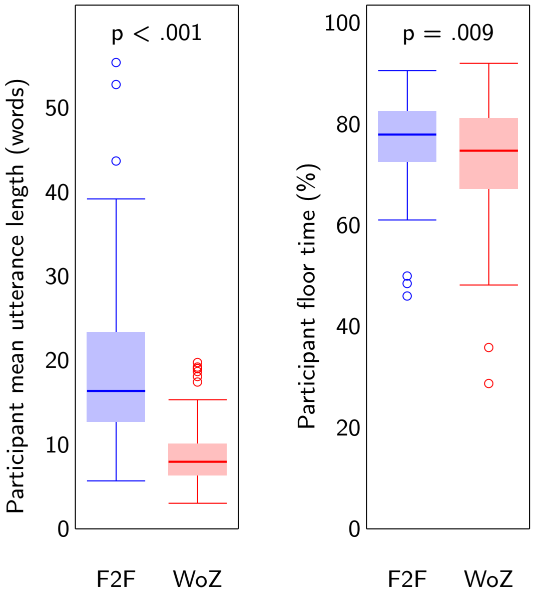


Figure 5: Utterance length and floor time.

in the Wizard-of-Oz dialogues. The wizard system did have verbal backchannels built in, but it was discovered during preliminary testing that participants tended to interpret these as an attempt by the interviewer to take the floor, and would subsequently stop speaking. As a consequence, the wizards did not use verbal backchannels during the main data collection, but only non-verbal backchannels. The verbal backchannels given by human interviewers in the face-to-face condition, in particular their ability to give specific feedback (Bavelas et al., 2000), may be a contributing factor which encourages longer participant utterances.

Participants also held the floor longer in the face-to-face condition, calculated as the proportion of participant speech duration out of total speech duration (Figure 5, right panel): median 77% of the total talking time, as compared to 75% in the Wizard-of-Oz condition; while the difference is not large, it is statistically significant ($W = 5261$, $p = 0.009$).[2]

There were also differences in speech quantity between distressed and non-distressed individuals, but only in the face-to-face condition (interaction between interview and distress conditions in a 2×2 ANOVA: $F(1,253) = 20$ for participant words, $F(1,253) = 15$ for participant time, $p < .001$ for both measures). The reason for this difference is the interview protocol: in the face-to-face dialogues, interviewers knew the participants' distress condition prior to the interview, and the protocol for interviewing distressed participants included more questions than for non-distressed participants. In the Wizard-of-Oz condition, wizards did not have access to the participants' medical condition, so the protocol was the same and there were no ensuing differences in dialogue length.

3.2 Filled pauses

Individuals in the Wizard-of-Oz condition produced filled pauses (*uh, um, mm*) at a rate almost twice that of individuals in the face-to-face condition: median 46 per thousand words in the wizard condition, 26 in the face-to-face condition ($W = 2556$, $p < .001$, Figure 6). The rate of utterances beginning with a filled pause was also significantly greater in the Wizard-of-Oz condition (median 19%) than in the face-to-face condition (median 9%; $W = 2580$, $p < .001$), possibly indicating that participants hesitated more when responding to the virtual interviewer. These findings are opposite to what is described in the literature, where people produce fewer disfluencies when talking to a voice-only, task-oriented dialogue system (Oviatt, 1995; Shriberg, 1996).

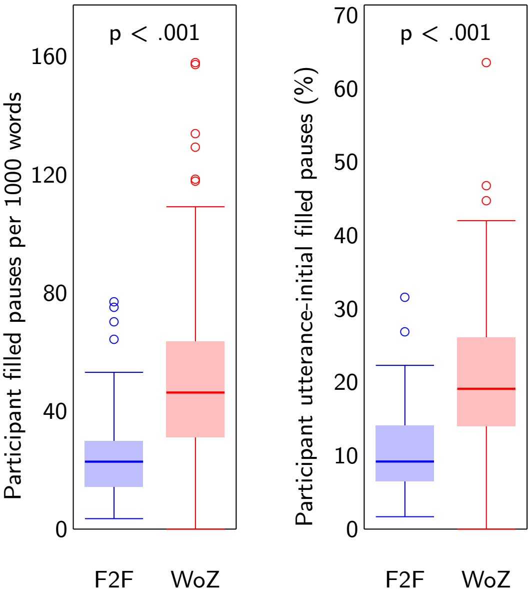


Figure 6: Filled pauses.

Even more striking is the relation between filled pause rate and utterance length. While previous literature has reported that longer utterances have higher rates of disfluency (Oviatt, 1995; Oviatt, 1996; Oviatt, 2000), our dialogues show the opposite: longer utterances have lower rates of filled pauses (Figure 7). The drop is rather dramatic, starting with the one-word utterances – 38% of these in the Wizard-of-Oz dialogues and 19% in the face-to-face dialogues consist of just a filled pause. The difference between the observed pattern and the one noted in previous literature is a further indication that the current dialogues are of a different nature than the ones investigated in the prior work.

We did not find a significant difference between the filled pause rates of distressed and non-distressed individuals. Working on a portion of the same data (43 dialogues from the second round of Wizard-of-Oz testing), DeVault et al. (2013) did find a significant difference, whereby distressed individuals produced fewer filled pauses per utterance than non-distressed individuals. This discrepancy is due to the fact that the current study uses more data, and employs a different depen-

---

[2] We excluded 3 Wizard-of-Oz dialogues from this test because errors in logging precluded the calculation of character speech time.

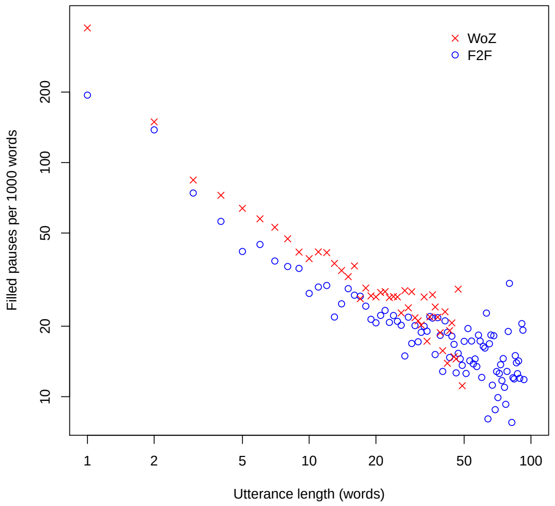


Figure 7: Relation between utterance length and filled pause rate. Data are plotted only when the corpus contains at least 10 utterances of a specified length.

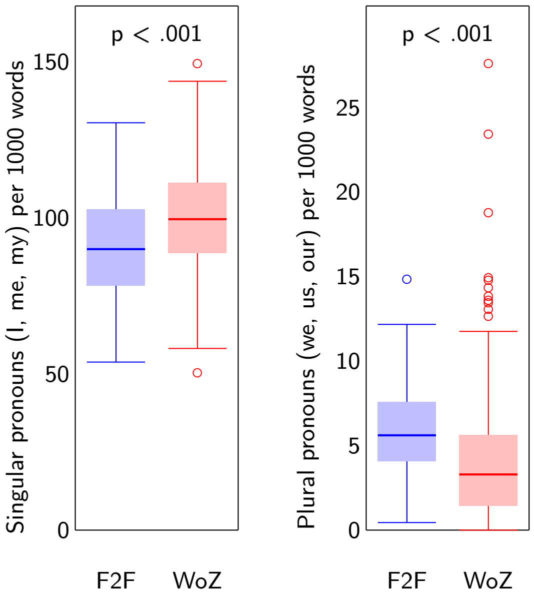


Figure 8: First-person pronouns

dent measure of disfluency (filled pauses per 1000 words rather than filled pauses per utterance). Measuring filled pauses per utterance on the full set of Wizard-of-Oz data failed to find a significant difference between distressed and non-distressed individuals, nor was a significant difference found when measuring filled pauses per 1000 words on the 43-dialogue subset.

3.3 Lexical items

An increased use of first-person singular pronouns has been linked to psychological distress in studies that compared the writing of suicidal and non-suicidal poets (Stirman and Pennebaker, 2001) and reflective essays by students (Rude et al., 2004); we tested these variables in order to see if these results carry over to dialogue. We did not find differences between distressed and non-distressed individuals or interactions between interview condition and distress condition, but we did find differences between the face-to-face and Wizard-of-Oz dialogues: first person singular pronouns (*I, me, my*) were used at a higher rate in the Wizard-of-Oz condition (median 100 per thousand words compared to 90 in the face-to-face condition, $W = 4608$, $p < .001$), whereas first-person plural pronouns (*we, us, our*) were used at a higher rate in the face-to-face condition (median 6 per thousand words compared to 3 in the Wizard-of-Oz condition, $W = 4075$, $p < .001$; see Figure 8). We do not have an explanation for these differences, and we cannot say whether they reflect a general difference between human-human and human-character interactions, or if they are caused by specific properties of the experimental setup in the two conditions. A sampling of the plural pronouns showed that they are primarily exclusive, that is they refer to the speaker and someone else but not the interviewer.

Face-to-face And uh I hooked up with uh somebody who runs this company uh at a party and uh we started talking and uh he offered me the job.

Wizard-of-Oz I have a stepfather and a half-brother we get along okay but we're not very close.

Other uses of *we* were generic.

Face-to-face And I'm a passionate believer in our trying to get our country going straight. I think we're we're going the wrong way and I don't know there's any way to stop it.

Wizard-of-Oz That's one of the things that took me a number of years to master though were my relaxation skills, I think that's a key thing and I think as we mature, as we learn how to do that, I wish I'd learned how to do that.

However, at least one participant referred to the virtual interviewer Ellie with an inclusive *we*:

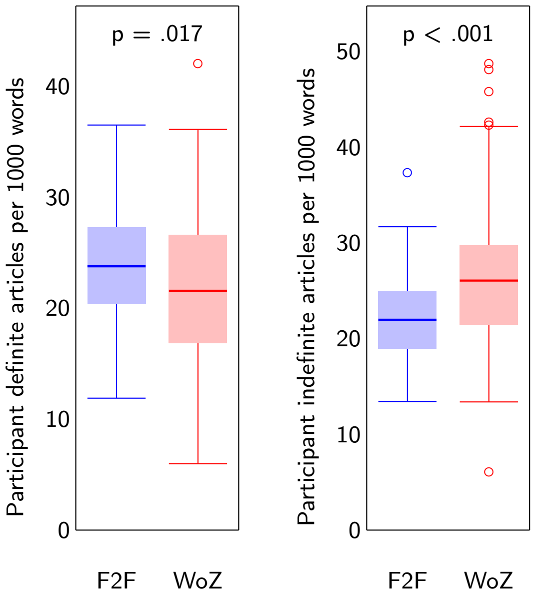


Figure 9: Articles.

Wizard-of-Oz Well, in the last few minutes since we started talking about depressing stuff, I starting to feel a little more down.

We also found a difference in the use of articles between the face-to-face and Wizard-of-Oz conditions: face-to-face dialogues contained more definite articles than Wizard-of-Oz dialogues (medians 24 and 22 per thousand words, $W = 5481$, $p = .02$), whereas the opposite is true for indefinite articles (medians 22 and 26 per thousand words, $W = 4263$, $p < .001$; Figure 9).

4 Discussion

Two main findings emerge from the present study. One is that human interviewers are able to engage participants in much longer conversations than Wizard-of-Oz characters. This is not surprising, given that the animated character has only a fixed number of utterances. Even in the short dialogue samples in Figure 1 above we can see how the human interviewer can tailor follow-up utterances to the participant's contributions, while the wizard-controlled character can only use generic follow-ups and has to move on when these are exhausted.

The second finding is that participants produce more filled pauses when talking to the animated interviewer than when talking to a human interviewer. This finding is important because it is the opposite of earlier results about computer-directed speech. Of course, the earlier results are from a very different kind of dialogue system – a disembodied, task-oriented dialogue interface as opposed to an animated conversational character. Nevertheless, these results have been taken to apply to computer-directed speech in general (e.g. Corley and Stewart, 2008, page 591: “Speakers tend to be more disfluent overall when addressing other humans than when addressing machines,” making reference to Oviatt, 1995). The present study shows that the results from disembodied task-oriented systems do not carry over to conversational dialogue systems, and more generally that computer-directed speech is not a unitary phenomenon, but that it varies depending on the computer system that the speech is directed to.

As mentioned in section 2.1, the face-to-face and Wizard-of-Oz dialogues were collected with the eventual goal of creating a fully automated character capable of interviewing people about mental distress. Experiments with an automated prototype are currently underway, and we hope to have access to dialogues between people and a fully automated character soon. Having a corpus with three types of comparable interview dialogue – human, human-controlled, and automated interviewers – will hopefully shed additional light on the question of the characteristics of computer-directed speech.

Acknowledgments

The first author was supported by the National Science Foundation under grant 1263386, “REU Site: Research in Interactive Virtual Experiences” (PI: Evan Suma).

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific under Contract N66001-11-C-4006, DARPA contract W911NF-04-D-0005, and the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, December.

Matthew Black, Jeannette Chang, Jonathan Chang, and Shrikanth S. Narayanan. 2009. Comparison of child-human and child-computer interactions based on manual annotations. In *Proceedings of the Workshop on Child, Computer and Interaction*, pages 2065–2068, Cambridge, MA, November.

Edward B. Blanchard, Jacqueline Jones-Alexander, Todd C. Buckley, and Catherine A. Forneris. 1996. Psychometric properties of the PTSD checklist (PCL). *Behaviour Research and Therapy*, 34(8):669–673, August.

Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147, June.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media / multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2065–2068, Lisbon, Portugal, May.

Justine Cassell, Kathleen Geraghty, Berto Gonzalez, and John Borland. 2009. Modeling culturally authentic style shifting with virtual peers. In *ICMI-MLMI '09: Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 135–142, Cambridge, Massachusetts, November. ACM.

Justine Cassell. 2004. Towards a model of technology and literacy development: Story listening systems. *Journal of Applied Developmental Psychology*, 25(1):75–105.

Martin Corley and Oliver W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of *um*. *Language and Linguistics Compass*, 2(4):589–602.

David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert (Skip) Rizzo, and Louis-Philippe Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*, pages 193–202, Metz, France, August. Association for Computational Linguistics.

Arne Jönsson and Nils Dahlbäck. 1988. Talking to a computer is not like talking to your best friend. In Thore Danielsen, editor, *Proceedings of The First Scandinivian Conference on Artificial Intelligence*, pages 53–68, Tromsø, Norway, March.

Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide – design and evaluation of a real-world application. In Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist, editors, *Intelligent Virtual Agents: 5th International Conference, IVA 2005, Kos, Greece, September 12–14, 2005 Proceedings*, volume 3661 of *Lecture Notes in Artificial Intelligence*, pages 329–343, Heidelberg, September. Springer.

Kurt Kroenke and Robert L. Spitzer. 2002. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9):509–515, September.

Sharon Oviatt. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Language*, 9(1):19–35.

Sharon Oviatt. 1996. Multimodal interfaces for dynamic interactive maps. In *Conference on Human Factors in Computing Systems: Common Ground (CHI '96)*, pages 95–102, Vancouver, BC, Canada, April. ACM.

Sharon Oviatt. 2000. Talking to thimble jellies: Children's conversational speech with animated characters. In *Sixth International Conference on Spoken Language Processing, (ICSLP 2000)*, volume 3, pages 877–880, Beijing, China, October. ISCA.

Stephanie S. Rude, Eva-Maria Gortner, and James W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert (Skip) Rizzo, and Louis-Philippe Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *10th IEEE International Conference on Automatic Face and Gesture Recognition*, Shanghai, China, April.

Elizabeth Shriberg. 1996. Disfluencies in Switchboard. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of ICSLP 96, the Fourth International Conference on Spoken Language Processing*, volume Addendum, pages 11–14, Philadelphia, October.

Lee Sproull, Mani Subramani, Sara Kiesler, Janet H. Walker, and Keith Waters. 1996. When the interface is a face. *Human-Computer Interaction*, 11(2):97–124, June.

Shannon Wiltsey Stirman and James W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4):517–522.

David Traum, William Swartout, Jonathan Gratch, and Stacy Marsella. 2008. A virtual human dialogue model for non-team interaction. In Laila Dybkjær and Wolfgang Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*, chapter 3, pages 45–67. Springer, Dordrecht.

Serdar Yildirim and Shrikanth Narayanan. 2009. Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):2–12, January.

On the use of social signal for reward shaping in reinforcement learning for dialogue management

Emmanuel Ferreira
LIA - University of Avignon
BP1228 - 84911 Avignon Cedex 9
Avignon, France
emmanuel.ferreira@univ-avignon.fr

Fabrice Lefèvre
LIA - University of Avignon
BP1228 - 84911 Avignon Cedex 9
Avignon, France
fabrice.lefevre@univ-avignon.fr

Abstract

This paper investigates the conditions under which social signals (facial expressions, postures, gazes, etc.), especially non-verbal multimodal user appraisal, can help to accelerate the learning capacity of a Reinforcement Learning (RL) agent in the dialogue management context. For this purpose a potential-based shaping reward method is used jointly with the Kalman Temporal Differences (KTD) framework so as to properly integrate the social aspects in an efficient optimization procedure through social-based additional reinforcement signals. Besides its general interest, this procedure could leverage system's development by allowing the designer to teach its system through explicit signals at its early stage of training. Experiments carried out using the state-of-the-art goal-oriented Hidden Information State (HIS) dialogue management framework in a simulation setup confirm the interest of the proposed approach.

.

1 Introduction

Goal-oriented statistical Spoken Dialogue Systems (SDSs), or even more generally Multimodal Dialogue Systems (MDSs), are the targets of this work. These systems are designed to achieve a task most often related to an information retrieval problem in collaboration with a human user (e.g. flight booking or hotel reservation services). The fundamental characteristic of this kind of "human-computer interface" is that the interaction between the human and the artificial agent (e.g. computer, robot, etc.) is mostly dominated by natural means of human communication (e.g. speech, gazes, gestures). The Dialogue Manager (DM) is the core component of SDSs, in charge of the interaction's course. It should infer the best decision sequence to fulfil the user goal. The dialogue management problem has first been described as a Markov Decision Process (MDP) in (Levin et al., 1997) and the Reinforcement Learning (RL) paradigm (Sutton and Barto, 1998) is employed to determine an optimal mapping between situations and actions, the policy. In this scheme the DM can be seen as an agent which has to interact with its environment (i.e. the user) in order to maximise some expected cumulative discounted reward. In most works the latter represents objective design criteria based on task completion and overall system efficiency. More recently, the MDP mathematical framework scheme was extended to Partially Observable Markov Decision Process (POMDP) to better cope with the inherent uncertainty on the information conveyed inside SDSs. This uncertainty comes from the fact that available pieces of information, collected from the user during consecutive dialogue turns, are extracted by error-prone input modules (e.g. speech recognizer, natural language understanding module, gesture recognizer, etc.). RL approaches were also successfully applied in this context (Young et al., 2010; Thomson and Young, 2010).

When developing a new SDS from scratch, in-domain dialogue corpora are seldom readily available and collecting such data is both time consuming and expensive (e.g. Wizard-of-Oz, prototyping). That is why, the capacity of a RL algorithm to learn online while interacting with the user is highly valuable. However, common approaches assume that an acceptable sub-optimal initial policy has been found by either exploiting user simulation methods (Schatzmann et al., 2005), or by hand (handcrafted dialogue manager) before any trials are made with real users. Recent works attempted to address this problem by using sample-efficient algorithms in order to limit the need of such a "bootstrap step". Thereby, TD-based SARSA with Gaussian Process (Gašić et al., 2010), incremental sparse Bayesian method (Lee and Eskenazi, 2012), or KTD (Daubigney et al., 2012) are among the most promising approaches. Anyhow, lowering the length of the warm-up learning phase, when the system can not interact with real users due to a high level of exploration and poor performance, is still an open problem when such systems are to be declined to real-world applications. One solution can be to introduce some initial expert knowledge (Williams, 2008) or to find ways to collect more hints from the environment which will accelerate the policy learning. For that purpose, we claim that social signals (Vinciarelli et al., 2009) can be employed as additional reinforcement signals (i.e. rewards) to refine and accelerate the policy optimization of a learning agent. Indeed, detecting social signals and social behaviours (e.g. emotions, turn taking attempts, politeness, noddings, postures, gazes, etc.) influence our everyday life behaviour in many ways (Custers and Aarts, 2005). Furthermore, by the fact that they can be gathered all along the dialogue, they may introduce a more granular view of the real quality of an interaction. Despite that some attempts to use emotion with RL have already been made (Broekens and Haazebroek, 2007), little has been done in the goal-oriented DM problem context. In this paper we propose a potential-based shaping reward method (Ng et al., 1999) to integrate these social aspects in combination with the use of the unified KTD framework with regards to its interesting properties (Geist and Pietquin, 2010; Daubigney et al., 2012). This preliminary study is carried out in a simulation setting where social reinforcement signals are simulated based on dialogue progress objective features representing the positiveness/negativeness of a particular situation. In this context, a better control over the experimental conditions, such as the simulated concept error rate level, is possible and comparison between several techniques is facilitated.

The remainder of the paper is organised as follows. In Section 2 some backgrounds on MDP/POMDP, RL paradigm, DM problem and KTD method are given. Then, in Section 3 social reward principle is detailed. Section 4 is dedicated to present the experimental setup. Then the following section details and comments on the different results obtained. Section 6 discusses on some considerations relevant to the use of social reinforcement, before concluding in Section 7 with some perspectives.

2 Background

This section briefly reviews the Markov Decision Processes (MDP) and the RL paradigm. Then, the casting of the DM problem as an MDP (POMDP) is presented. Finally, the KTD method is concisely introduced.

2.1 Markov Decision Processes

A tuple $\{S, A, T, R, \gamma\}$ forms a MDP, where S is the state space (discrete, continuous or mixed), A is the discrete action space, T is a set of Markovian transition probabilities, R is the immediate reward function, $R : S \times A \times S \rightarrow \Re$ and $\gamma \in [0, 1]$ the discount factor (discounting long term rewards). The environment evolves at each time step t to a state s_t and the agent picks an action a_t according to a policy mapping states to actions, $\pi : S \rightarrow A$. Then state changes to s_{t+1} according to the Markovian transition probability $s_{t+1} \sim T(.|s_t, a_t)$ and, following this, the agent received a reward $r_t = R(s_t, a_t, s_{t+1})$ from the environment. The overall problem of MDP is to derive an optimal policy maximising the reward expectation. Typically the averaged discounted sum over a potentially infinite horizon is used, $\sum_{t=0}^{\infty} \gamma^t r_t$. Thus, for a given policy and start state s, this quantity is called the value function: $V^{\pi}(s) = E[\sum_{t \geq 0} \gamma^t r_t | s_0 = s, \pi] \in \Re^S$. V^* corresponds to the value function of any optimal policy π^*. The Q-function may be defined as an alternative to the value function. It adds a degree of freedom on the first selected action, $Q^{\pi}(s, a) = E[\sum_{t \geq 0} \gamma^t r_t | s_0 = s, a_0 = a, \pi] \in \Re^{S \times A}$. As well as V^*, Q^* corresponds to the action-value function of any optimal policy π^*. If it is known, an optimal policy can be directly computed by being greedy according to Q^* , $\pi^*(s) = \arg\max_a Q^*(s, a) \forall s \in S$.

2.2 Dialogue Management as a POMDP

Dialogue management problem has first been described in (Levin et al., 1997) as a Markov Decision Process to determine an optimal mapping between situations and actions. The POMDP framework (Kaelbling et al., 1998), as a generalization of the fully-observable MDP, maintains a belief distribution $b(s)$ over user states, assuming the true one is unobservable. Thereby, POMDP explicitly handles parts of the inherent uncertainty of the DM problem (e.g. word error rate, concept error rate). A POMDP policy maps the belief state space into the action space. That is why, the optimal policy can be understood as the solution of a continuous space MDP. In practice, POMDP problems are intractable to solve exactly due to the curse of dimensionality (i.e. belief state/action spaces). Among other techniques, the HIS model (Young et al., 2010) circumvents the RL scaling problem by organising the belief space into partitions, grouping states sharing the same probability, and then mapping the full belief space (partitions) into a much reduced summary space where RL algorithms work reasonably well.

Although variants have been proposed and tested, e.g. (Pinault and Lefèvre, 2011), HIS remains a reference. However, the choice of a Monte Carlo Control RL algorithm (Sutton and Barto, 1998) is still questioned and recent studies show the interest of considering sample-efficient algorithms for the DM problem (Gašić et al., 2010; Daubigney et al., 2012). More especially (Daubigney et al., 2012) showed that KTD framework offers a unified framework able to cope with all DM required properties: it is sample-efficient, it allows on-policy/off-policy learning through two algorithms (respectively KTD-Q and KTD-SARSA) which can both perform online and offline learning, it provides ways to deal with the "exploration/exploitation" dilemma using uncertainty on value estimate, it allows value tracking, and it supports linear and non-linear parametrisation. Furthermore, KTD algorithms were favourably compared to different state-of-the-art algorithms able to deal with one single property at once, such as Q-learning, LSPI or GP-SARSA.

2.3 The KTD Framework

The Kalman Temporal Differences (KTD) framework (Geist and Pietquin, 2010) is derived from the well-known Kalman filter algorithm (Kalman, 1960) aiming at inferring some hidden variables from related past observations and applied to the estimation of the temporal differences for the action-value function optimisation. In this framework, a parametric representation of the Q-function is chosen: $\hat{Q}_\theta = \theta^T \phi(s,a)$, where the feature vector $\phi(s,a)$ is a set of n basis functions to be designed by the practitioner and $\theta \in \Re^n$ the parameter vector to be learnt. Notice that just the very basic explanations are recalled here, for further details please refer to (Geist and Pietquin, 2010; Daubigney et al., 2012). The components of the parameter vector θ are the hidden variables which are modelled as a random vector. Such parameter vector is considered to evolve following a random walk though this evolution equation: $\theta_t = \theta_{t-1} + v_t$, with v_t a white noise of covariance matrix P_{v_t}. The latter allows to take into account the possible non-stationarity of the function. The observations correspond to the environment rewards which are linked to the hidden parameter vector through one of the sampled Bellman equations $g_t(\theta_t)$ depending on the RL scheme employed (i.e. evaluation for on-policy or optimality for off-policy learning):

$$g_t(\theta_t) = \begin{cases} \hat{Q}_{\theta_t}(s_t, a_t) - \gamma \hat{Q}_{\theta_t}(s_{t+1}, a_{t+1}) \\ (evaluation) \\ \hat{Q}_{\theta_t}(s_t, a_t) - \gamma \max_a \hat{Q}_{\theta_t}(s_{t+1}, a) \\ (optimality) \end{cases}$$

Rewards are supposed to follow the observation equation: $r_t = g_t(\theta_t) + n_t$ where a white noise n_t with covariance matrix P_{n_t} is also considered. Two algorithms can be defined: KTD-SARSA which denotes the use of the sampled evaluation Bellman equation and KTD-Q, the use of the sampled optimality one.

3 Social Reinforcement

In this section a rather simple definition of social reward is given followed by a mathematical formalisation of such a reward. Then, a method to simulate social signal is described.

3.1 Definition and formalisation

Social signal is a generic term which encompasses all the behavioural cues which can be encountered during an interaction with a human (e.g. blinks, smiles, crossed arms, laughter, nodding and the like). Social RL consists hence in exploiting these cues in order to guide the learning process. However, the agent can use this information in multiple ways: as reinforcement, as additional information integrated into the user state or as meta-parameter (e.g. in an exploration/exploitation scheme). Moreover, one may also think of using emotion in the system response (*emotional agent*) and thus, make use of this information so as to improve its own social behaviour.

This work focuses on extracting social rewards based on positive and negative social signals emitted by the user and use them as additional rewards (or punishments). At each dialogue turn, a social reward may be perceived by the system. In this scenario a social signal can be seen as a user behaviour attesting its own judgement on the state evolution. Ergo, the social reward corresponds to the associated positiveness or negativeness of this signal represented as a signed real value. In that purpose we propose to consider the social reward function as a shaping reward function. The memoryless shaping reward function, which is one of the most general shaping pattern, is adopted here. So, the considered reward function is the sum of the basic environment reward function R_{env} (objective) and the new social one R_{social} (subjective). The resulting transformed MDP $M^{'}$ is defined by the tuple $(S, A, T, \gamma, R^{'})$ where $R^{'}$ is is the reward function defined as: $R^{'}(s_t, a_t, s_{t+1}) = R_{env}(s_t, a_t, s_{t+1}) + R_{social}(s_t, a_t, s_{t+1})$ where $R_{social} : S \times A \times S \rightarrow \Re$ is a bounded real-valued function called here the social-shaping reward function. Since the system is learning a policy for $M^{'}$ in the idea of using it in M, the question at hand is: what form of social-shaping reward function R_{social} can guarantee that the optimal policy in $M^{'}$ will be optimal in M? In the case where no further knowledge of T and R dynamics is available (no expert), a potential-based shaping reward leave (near-)optimal policies unchanged (Ng et al., 1999). Thereby, the potential-based shaping reward function is adopted for R_{social}, corresponding to function F in Ng et al.'s paper, and can be defined as follows:

$$R_{social}(s_t, a, s_{t+1}) = \gamma\psi(s_{t+1}) - \psi(s_t) \qquad (1)$$

where ψ is a potential function, here computed using a heuristic score based on the social signal.

3.2 Social agenda-based simulation

3.2.1 Goal and agenda-based simulation

As indicated before, the social rewards are simulated in this study. We used the agenda-based user simulator presented in (Schatzmann et al., 2006) wherein the user is simulated at the intentional semantic level (i.e. dialogue act level). This approach factors the user state into an agenda A and a goal G: $S = (A, G)$, where $G = (C, R)$. The goal G ensures that the simulated user reacts in an appropriate, consistent and goal-oriented manner. It consists of a set of constraints C specifying the required properties that the system should satisfy (they are the objects of the negotiation) and a set of requests R which represent the desired pieces of information (e.g. address, phone number, available schedules). The agenda A is a stack-like structure containing the pending user acts that are deemed necessary to elicit the information specified in the goal. For further details on this simulation method please refer to both (Schatzmann et al., 2006) and (Keizer et al., 2010).

$$
\begin{array}{lcl}
C_0 & = & \left[\begin{array}{l}\text{type=bar}\\ \text{drinks=beer}\\ \text{area=east}\end{array}\right]\\
R_0 & = & \left[\begin{array}{l}\text{name=}\\ \text{addr=}\\ \text{phone=}\end{array}\right]\\
Sys_1 & & \text{Hello, how may I help you ?}\\
A_1 & = & \left[\begin{array}{l}\text{inform(type=bar)}\\ \text{inform(drinks=beer)}\\ \text{inform(area=east)}\\ \text{request(name)}\\ \text{request(addr)}\\ \text{request(phone)}\\ \text{bye()}\end{array}\right]\\
Usr_1 & & \text{I'm looking for a nice bar serving beer}\\
Sys_2 & & \text{Ok, a wine bar. What part of town do}\\
 & & \text{you have in mind ?}\\
A_2 & = & \left[\begin{array}{l}\text{negate(drinks=beer)}\\ \text{inform(area=east)}\\ \text{request(name)}\\ \text{request(addr)}\\ \text{request(phone)}\\ \text{bye()}\end{array}\right]\\
Usr_2 & & \text{No, beer please !}\\
Sys_3 & & \text{A bar serving beer, correct ?}\\
A_3 & = & \left[\begin{array}{l}\text{affirm()}\\ \text{inform(area=east)}\\ \text{request(name)}\\ \text{request(addr)}\\ \text{request(phone)}\\ \text{bye()}\end{array}\right]\\
Usr_3 & & \text{Yes something in the east}\\
 & & \text{part of town please !}\\
Sys_4 & & \text{There is no bar serving beer in the east}\\
 & & \text{part of town. But the Kings Shilling, near}\\
 & & \text{the Castle, is a very nice bar serving beer.}\\
C_4 & = & \left[\begin{array}{l}\text{type=bar}\\ \text{drinks=beer}\end{array}\right]\\
R_4 & = & \left[\begin{array}{l}\text{name=Kings Shilling}\\ \text{addr=Castle Loop}\\ \text{phone=}\end{array}\right]\\
A_4 & = & \left[\begin{array}{l}\text{request(phone)}\\ \text{bye()}\end{array}\right]\\
Usr_4 & & \text{Ok, That is just the way it is. What's}\\
 & & \text{the phone number ?}\\
Sys_5 & & \text{Could you please rephrase that ?}\\
A_5 & = & \left[\begin{array}{l}\text{request(phone)}\\ \text{bye()}\end{array}\right]\\
Usr_5 & & \text{What's the phone number ?!}\\
Sys_6 & & \text{The number is 728-64-32.}\\
R_6 & = & \left[\begin{array}{l}\text{name=Kings Shilling}\\ \text{addr=Castle Loop}\\ \text{phone=728-64-32}\end{array}\right]\\
A_6 & = & \left[\begin{array}{l}\text{bye()}\end{array}\right]\\
Usr_6 & & \text{Thanks, goodbye!}
\end{array}
$$

Figure 1: Sample dialogue and its corresponding agenda sequence

Figure 1 illustrates how the simulation works and how tracking both the agenda and the goal evolution can serve to detect some situations when social signals are prone to be generated. For instance, in A_2, the presence of a *negate* act at the top of the agenda means that a user constraint has been violated (here `drinks=beer`). So, it is a negative cue. In the same way, the `affirm` act in A_3 underlines a positive situation. That is why, the nature of the top dialogue act of the agenda can give an insight into the positiveness or the negativeness of the user state evolution.

3.2.2 Social cues

Table 1 presents some simple positive and negative cues extracted from the agenda and goal structures in the user simulator during dialogue simulations. Each of them is weighted in order to give more or less emphasis on specific features. Although a continuous scale is possible, a five-point agreement scale (Likert scale) is adopted here for ψ with regard to the way subjective measures are gathered in PARADISE (Walker et al., 1997). Each level is associated with a representative real number associated with an agreement scale, from strongly negative $(--)$ to strongly positive $(++)$. So, after a normalisation step the sum of all the simulated social features gives an overall score C_{s_t} which is rescaled on a five-point Likert scale using a threshold ξ. Thus, at each time step t, a "potential-like" social reward is computed using Eq 1 and ψ function:

$$\psi(s) = \begin{cases} -1 & ,if\ C_s < -\xi & (--) \\ -0.5 & ,if\ -\xi \leq C_s < 0 & (-) \\ 0 & ,if\ C_s = 0 & (neutral) \\ 0.5 & ,if\ 0 < C_s \leq \xi & (+) \\ 1 & ,if\ C_s > \xi & (++) \end{cases}$$

Positive Cues		Negative Cues	
1	Positive top dialogue act type (e.g. affirm, confirm)	1	Negative top dialogue act type (negate, deny, etc.)
2	Number of slots filled	2	Agenda size
3	Partial completion flag	3	Dialogue length
4	Final completion flag	4	Top agenda act contains already transmitted item

Table 1: List of positive and negative cues collected from agenda and goal

s_t	Positive cues	Negative cues	$\psi(s_t)$	R_{social}
s_3	1(1)	2(-6), 3(-4)	0.5	0.45
s_4	2(2/3) 3(1)	2(-2) 3(-5)	1	

Table 2: Social reward computation example

The process of social reinforcement reward computation can be decomposed into two steps. First, the gathering of positive and negative social cues from the factored user state. Second, the social reward estimation using the potential-based social reward function. An example of such a process is summarised in Table 2. The first column represents the analysed user state s_t (i.e. the corresponding agenda A_t and goal G_t in Fig. 1). The second and the third columns are respectively the lists of positive and negative cues which have been detected (using the id from Tab. 1) and their associated value in brackets. For example, in the first row and third column, cue 2 corresponds to the number of items in the agenda and the value 6 is extracted from A_3, minus sign indicates negativeness of the cue. The fourth column corresponds to the ψ value (i.e. the Likert score). It is computed applying some weights on the detected cue values. As an illustration, for the negative cue 3, $1/30$ is chosen as weight because the maximum number of turns allowed by the system is 30. Consequently, $1/30$ can be viewed as a normalisation value. It is important to notice that such weights have been determined following some expert intuitions. They have been chosen to correspond to an average user appraisal of the dialogue progress. In (Ferreira and Lefèvre, 2013), different user profiles are designed by varying these weights to study to what extent social signals can help user adaptation capacities of a learning agent. The last column shows the resulting social reward applying Eq. 1 with $\gamma = 0.95$. The positive score 0.45 denotes a quite favourable evolution between s_3 and s_4. To compete with the environment reward the social reward can be rescaled using an exponent. In real applications social cues could be elicited using several multimodal social detectors (e.g emotion face tracking, gesture classification, social keyword spotting). These latters may produce a list of detector-specific positive and negatives cues. For instance, the face tracker may produce a cue dedicating to smile detection which value is the probability of its inner model thereby consisting in a positive cue, likewise the definition of two lists of negative/positive keywords may help to produce two polarized cues from their detection in the ASR results associated with their posterior probabilities. Then, the same mecanism of a weighted interpolation could be used to infer $\psi(s)$ from the valued cues output by the various detectors.

4 Experimental Setup

First, the HIS-based Dialogue System is briefly described. Then, some details on experimental conditions are given.

4.1 TownInfo Dialogue System

The TownInfo Dialogue System (Young et al., 2010) is a HIS-based dialogue system for the tourist information domain, related to a virtual town. The TowInfo system has already been tested with real users in (Schatzmann et al., 2006), and in a more recent and matured version, called CamInfo (Cambridge tourist information), in (Gašić et al., 2010). In order to deal with large state and action space the system maintains a set of partitions which represent the overall belief state. Both the latter and the action space are mapped into more reduced summary spaces where RL algorithms are tractable. The summary state space is the compound of two continuous values (the two-first top partitions probabilities) and three discrete values (last user act type and a partition and a history status). The summary action space contains 11 actions (e.g. inform, confirm). The environment rewards penalised each dialogue turn by -1 and at the end of a dialogue the DM is rewarded a +20 bonus if the goal is reached, nil otherwise.

4.2 Experimental details

To assess the performance of introducing social cues as a reinforcement signal, the online version of the off-policy KTD-Q algorithm (noted KTD-Q BASELINE) is employed as our baseline due to its high performance in the conditions at hand (Daubigney et al., 2012). The Q-function is parametrised using linear-based Radial Basis Function (RBF) networks, one per action, as described in (Daubigney et al., 2012) and the Bonus-Greedy scheme (Daubigney et al., 2011) is adopted, with $\beta = 1000$ and $\beta_0 = 100$. The discount factor γ is set to 0.95 in all experiments. By default, the user simulator is set to interact with the DM at a 10% concept error rate. The weight coefficient of the overall social reward is set to 4 and $\xi = 0.3$, likewise all other individual cues are weighted manually. All the results are averaged over 50 independent training under online RL conditions and are presented in terms of mean discounted cumulative rewards with respect to both the number of training dialogues (i.e. samples) or different CER levels. The associated standard deviations are added to all the results. The authors consider that the average cumulative environment rewards can be sufficient metric to compare the different approaches. This is explained by the fact that in the environment reward function the success (full user goal completion) is rewarded by a +20 bonus and failure and elapsed time (turn) respectively punished by a 0 and -1. For comparison purposes all the experiments with a social reward presented in our plots are given in terms of the environment reward, R_{env}, only.

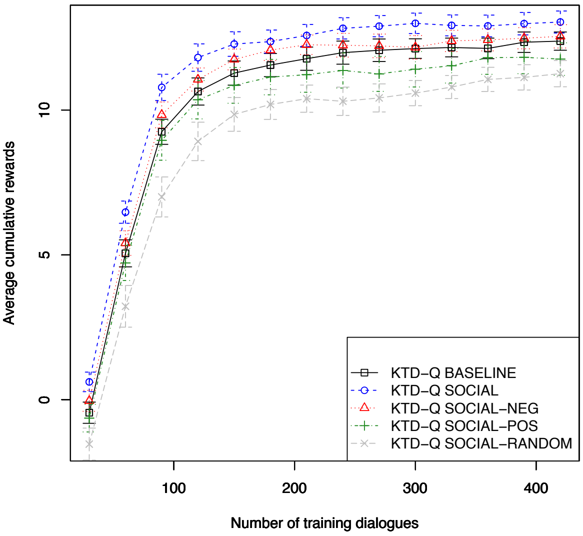


Figure 2: Results of 4 different configurations of the social-shaped KTD-Q algorithm compared to KTD-Q baseline during the learning of the policy (controlled case)

5 Results

This section presents the results obtained using the agenda-based user simulator described in Section 3.2.

5.1 Online policy using social reinforcement learning

In this section the benefits of adding social reinforcement signals for optimizing the DM policy are evaluated considering several social reinforcement configurations which take into account different kind of cues for the social reward computation. The classic approach noted KTD-Q SOCIAL considers both the negative and the positive social cues, as described in Section 3.2.2.

Results are shown in Figure 2 in terms of cumulative discounted environment rewards gathered during the learning stage of the policy (controlled case) when exploration is possible. For these curves, each point is an average of the 50 independent learning performance using a sliding window of 100 point width. Only the first 500 dialogues are considered here because we want to focus on the early stage of training for which system performance is critical. We can observe that KTD-Q SOCIAL slightly outperforms KTD-Q BASELINE in terms of both the final learned performance, which is better of about 0.5 turn on average, and the learning time to achieve a similar performance level, which is reduced. For example, the performance obtained performing 200 dialogues with KTD-Q BASLINE algorithm are reached at about 100 dialogues using KTD-Q SOCIAL. Furthermore, a comparison between three other kinds of configuration of the simulated social signal is also made. The first (KTD-Q SOCIAL-NEG) and the second (KTD-Q SOCIAL-POS) configurations are respectively using only the negative or positive social cues. The third configuration is a randomized social signal generator (KTD-Q SOCIAL RANDOM). As expected, KTD-Q SOCIAL-RANDOM is the worst, followed by KTD-Q SOCIAL-POS, KTD-Q BASELINE and KTD-Q SOCIAL-NEG. KTD-Q SOCIAL which combines both positive and negative cues still obtains the best results. All configurations (except KTD-Q SOCIAL-RANDOM) are rather close if we consider the confidence radius of their results. However an important point is that even in the case of random social reinforcement, the potential-based technique ensures that convergence to the near-optimal policy is still preserved. From this experiment it seems that the convergence is better guided by negative information which is an interesting finding considering that negative emotions might be easier to emit and detect in a real setup.

5.2 Online policy in noisy conditions

Eventually we intend to evaluate the impact of noise on the proposed optimization procedure. Noise robustness is studied in terms of CER, Environment and Social Reward Error Rates, noted respectively ERER, SRER. Although the previous experiment has shown encouraging results when social reinforcement is considered, it should be kept in mind that in the previous conditions social signals are perfectly perceived by the learning agent. In a more realistic setup like user trials such signals, due to their inherent complexity (e.g. multimodal aspects, context-dependent interpretation) cannot be perfectly observed. This difficulty is introduced in the simulation by means of an artificial SRER. At a given rate the social cues are randomly modified to the inverse of what they should be. In the same way, when online learning is adopted the user should mark the overall dialogue in terms of task completion (objective metric). But, as shown in (Gašić et al., 2010), the feedback given by a real user can be erroneous. This will be reflected by the ERER in our experiments. At a certain rate the final evaluation of dialogue success (correct or not) is inverted. Wrong feedbacks can be explained by the subjectivity of the task. Although the goal is achieved any inconsistent behaviour of the system during the dialogue can drive the user to penalise the system at the end, but also by the fact that a trial user is not really committed to the task, if the system fails there is no consequence for her or if the system asks for some constraint release the user has no personal rationale to guide her behaviour. In any case, the quality of the reward function is crucial for the RL algorithms as the speed of convergence to the optimal policy relies on it. In addition, the presence of high CER level also has a negative influence when this additional difficulty is present from the beginning of the learning (no progressive degradation).

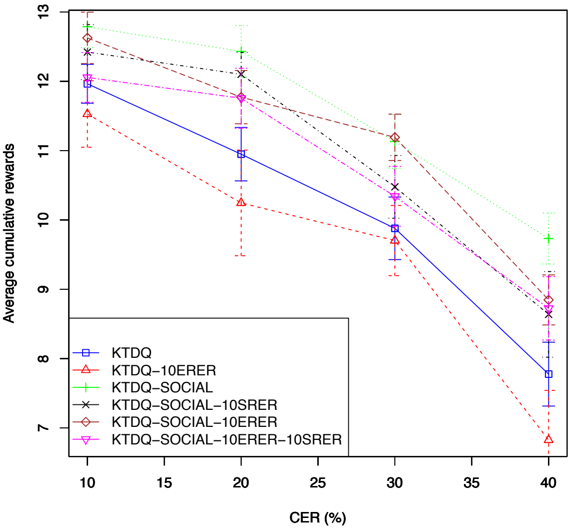


Figure 3: Results of baseline and social-shaped KTD-Q algorithms in different noise conditions (no control)

Here, 7 methods are compared: KTD-Q BASELINE and KTD-Q BASELINE-10ERER, KTD-Q SOCIAL, KTD-Q SOCIAL-10ERER, KTD-Q SOCIAL-10SRER and KTD-Q SOCIAL-10ERER-10SRER. The 10XER mean that the corresponding error rate X is set to 10%. Results are shown in Figure 3 in terms of cumulative rewards with respect to different CER levels. For these curves, each point is an average made over the results obtained using 50 policies learned with 400 dialogues and then tested with 1000 dialogues. In the latter test setup, the next action is cho-

Use social	SRER	Rewards	Success rate
no	-	10.24 (±0.76)	91.14 (±1.58)
yes	0	11.77 (±0.38)	93.42 (±0.80)
yes	10	11.75 (±0.43)	93.73 (±0.58)
yes	20	11.28 (±0.45)	92.53 (±0.88)
yes	30	10.80 (±0.42)	91.68 (±1.10)
yes	40	10.67 (±0.43)	91.33 (±1.01)
yes	50	10.06 (±0.71)	89.34 (±3.70)

Table 3: Results of KTD-Q algorithm at 20% CER and 10 % ERER using different SRER levels (no control)

sen greedily with respect to the learnt Q-function (no exploration). Considering only the KTD-Q BASELINE and KTD-Q-BASELINE-10ERER the influence of CER and ERER can be easily identified. Thus, as the ERER and the CER increase the overall performance decreases. Nevertheless, in all conditions the use of a social reinforcement has a positive impact on the performance of the KTD-Q algorithm. Thus, social reinforcement improves the ability to defer the impact of noise in terms of both CER and ERER. One of the reasons for this is that social rewards are gathered all along the dialogue and offer a granular form of reward function. So, in case of the user giving an erroneous final reward, collected positive and negative social rewards can counterbalance this mistake (as an hint of the overall user satisfaction). Furthermore, in case of high CER, social rewards can favour or penalize a system local behaviour despite the overall task failure (or success). However, the benefit of social reinforcement tends to decrease as the SRER raises. Thereby, in order to study the impact of SRER alone, Table 3 is populated with the results obtained with different SRER levels at 20 % CER and 10 % ERER, both corresponding to realistic values for field trials. Above 30% SRER, taking into account social signals seems to be unnecessary or even disadvantageous. Actually, even if the results obtained with 40% SRER are slightly better than those obtained with the baseline, they do not converge as quickly (e.g. considering 200 training dialogues the baseline outperforms this social version). It is worth noting that ERER and SRER are simulated with no specific prior assumption. Indeed a rather simple random error approach is used. In more a sophisticated framework, such errors could be learnt from data.

6 Discussions

In this “proof of concept” study a simulation setup has been adopted, but undeniably real user trials are required to validate the suggested claims presented all along. Mechanisms to extract correctly social signals through multimodal cues from real user have to be envisaged as for instance what is done in the INTERSPEECH Computational Paralinguistics Challenge (Schuller et al., 2012). Even if the capacity of these methods remains highly imperfect if these cues are gathered in an unconstrained and implicit manner (Vinciarelli et al., 2009), the experiments in Section 5.2 show that we can evaluate them with a certain level of imprecision without jeopardizing the merits of the proposed method. Furthermore, we assume that this problem can be simplified if we consider an interaction with a cooperative and rational “seed user” (e.g. a system designer), which employs a limited set of non-verbal cues (e.g. head gesture, tone) in order to accelerate the learning process. The use of social rewards allows a more granular view of the reward function rather than a binary judgement at the end of the episode. So, it serves as a more specific way to avoid or strengthen some local system behaviours. Thereby, when sample-efficient algorithms are considered the approach can be viewed as a way to avoid the need for a user simulator by using 100-200 interactions with a seed user to bootstrap the system performance. Such setup can be assimilated to active learning like what is done in (Doshi and Roy, 2008) and thus linked to imitation-based (Price and Boutilier, 2003) or inverse approaches to RL as in (Chandramohan et al., 2011).

7 Conclusion

This paper has described a method by which social based reinforcement learning can be used to train a dialogue policy from scratch in just a few hundred dialogues and that improves the baseline performance in terms of rapidity of convergence. The approach also shows better robustness to noisy conditions in terms of semantic input error rate and environment reward error rate. The presented method also has interesting properties that guarantee the optimality when social signals are merged into an additional reinforcement learning signal using an amenable potential-based shaping reward function to introduce the detected social cues as additional reinforcement signals. In the present work the social signals were simulated from an agenda-based user simulator and thus real user trials are still needed to uphold our claims.

8 Acknowledgments

The authors would like to thank the Cambridge University Dialogue Systems Group for providing the TownInfo HIS System. As well as Lucie Daubigney, Olivier Pietquin and the MALIS Supélec-Metz Group for their help in using the KTD Framework. This work is partially funded by the ANR MaRDi project.

References

Joost Broekens and Pascal Haazebroek. 2007. Emotion and reinforcement: Affective facial expressions facilitate robot learning. In *Artificial Intelligence for Human Computing*, volume 4451 of *Lecture Notes in Computer Science*, pages 113–132.

Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefèvre, and Olivier Pietquin. 2011. User Simulation in Dialogue Systems using Inverse Reinforcement Learning. In *Interspeech*.

Ruud Custers and Henk Aarts. 2005. Positive affect as implicit motivator: On the nonconscious operation of behavioral goals. *Journal of Personality and Social Psychology*, 89(2):129–142, August.

Lucie Daubigney, Milica Gašić, Senthilkumar Chandramohan, Matthieu Geist, Olivier Pietquin, and Steve Young. 2011. Uncertainty management for on-line optimisation of a pomdp-based large-scale spoken dialogue system. In *Interspeech*.

Lucie Daubigney, Matthieu Geist, Senthilkumar Chandramohan, and Olivier Pietquin. 2012. A comprehensive reinforcement learning framework for dialogue management optimization. *Journal on Selected Topics in Signal Processing*, 6(8):891–902.

Finale Doshi and Nicholas Roy. 2008. Spoken language interaction with model uncertainty: an adaptive human–robot interaction system. *Connection Science*, 20:299–318.

Emmanuel Ferreira and Fabrice Lefèvre. 2013. Social signal and user adaptation in reinforcement learning-based dialogue management. In *IJCAI 2nd Workshop on Machine Learning for Interactive Systems*.

Milica Gašić, Filip. Jurčíček, Simon Keizer, François Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *SIGDIAL*.

Matthieu Geist and Olivier Pietquin. 2010. Kalman temporal differences. *Journal of Artificial Intelligence Research (JAIR)*, 39(1):483–532, September.

Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence Journal*, 101(1-2):99–134, May.

Rudolf E. Kalman. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45.

Simon Keizer, Milica Gašić, Filip Jurčíček, François Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Parameter estimation for agenda-based user simulation. In *SIGDIAL*.

Sungjin Lee and M. Eskenazi. 2012. Incremental sparse bayesian method for online dialog strategy learning. *Journal on Selected Topics in Signal Processing*, 6:903–916.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 1997. Learning dialogue strategies within the markov decision process framework. In *ASRU*.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*.

Florian Pinault and Fabrice Lefèvre. 2011. Unsupervised clustering of probability distributions of semantic graphs for pomdp based spoken dialogue systems with summary space. In *IJCAI 7th Workshop on knowledge and reasoning in practical dialogue systems*.

Bob Price and Craig Boutilier. 2003. A bayesian approach to imitation in reinforcement learning. In *IJCAI*.

Jost Schatzmann, Matt Stuttle, Karl Weilhammer, and Steve Young. 2005. Effects of the user model on simulation-based learning of dialogue strategies. In *ASRU*.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review*, 21(2):97–126, June.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2012. Paralinguistics in speech and language - state-of-the-art and the challenge. *Computer Speech and Language (CSL), Special Issue on " Paralinguistics in Naturalistic Speech and Language"*, 27(1):4–39, Jan.

Richard S. Sutton and Andrew G. Barto. 1998. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054.

Blaise Thomson and Steve Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech and Language*, 24(4):562–588.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. Paradise: a framework for evaluating spoken dialogue agents. In *ACL*.

Jason D. Williams. 2008. Integrating expert knowledge into pomdp optimization for spoken dialog systems. In *Proceedings of the AAAI-08 Workshop on Advancements in POMDP Solvers*.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174.

When Apples Mean Oranges: Lexical De-Entrainment and Alignment Theories

Christina Freihorst, Simón Algorta, Vivien Mast, Robert Porzel
University of Bremen, Bibliothekstrae 1, 28359 Bremen
`ch.freihorst@uni-bremen.de`, `algorta@informatik.uni-bremen.de`,
`viv@tzi.de`, `porzel@tzi.de`

Abstract

This paper investigates the effects of semantic distance on the development of lexical entrainment. For this purpose, the authors developed a card game with three levels of semantic distance. The participants were asked to arrange the cards into a congruent sequential order. By increasing the semantic distance, more words were needed to solve the task and a higher rate of hypernyms was used, demonstrating lexical entrainment. Additionally, results showed that the participants recurred to the use of de-entrained terms on a third stage of the conversation. Based on this we examine what this finding might entail for existing theories on linguistic alignment.

1 Introduction

Referring to objects is a central part of human communication. It is well known that we do not only learn that every object has its name, but also that an object's name is not invariant – in different situations one can refer to the same object as 'tree', 'oak' or 'plant' depending on the context (Hermann & Deutsch, 1976; Furnas, Landauer & Dumais, 1987). The context may hinge on the range of objects from which the referent needs to be distinguished or on social aspects of a situation (Herrmann & Grabowski, 1994). This variability in reference underlies certain principles, such as avoidance of confusion and gauging the expectations of the listener.

A process that plays a central role in reference resolution is lexical entrainment: During the course of a conversation, interlocutors show a tendency to converge on a common set of referring expressions.

In order to examine the processes that underlie reference in a dialogue situation that involves hypernymy[1] and hyponymy[2], we examined how participants deal with the challenge of finding appropriate referring expressions when aligning semantically related but not identical concepts. For this purpose we designed a card game with three levels of increasing semantic distance, where semantic distance increases when the distance between hypernyms and hyponyms increases as defined by the number of steps needed to traverse the WordNet graph from one to the other (Fellbaum, 1994). The two participants got different sets of cards where each card of one participant corresponded to a semantically related card of the other. Participants were not able to see the cards of their partner. The exercise was to order the cards in congruent order, for this purpose the participants had to refer to the cards, developing a strategy for bridging the semantic gap between the corresponding cards. This means, they had to come to understand that when one talks about one referent (e.g. apple), the other needs to consider the related referent within his own set of cards (e.g. orange). Our primary aim was to discover how lexical entrainment changes with increasing semantic distance of the objects that need to be referred to. We assumed that entrainment would become more difficult and take longer with increasing semantic difference between the objects.

In the following sections, we will describe cognitive accounts of reference and lexical entrainment, focusing on the controversial aspects of different theories of entrainment. Then we will present the experimental setup and the results of the experiment. Finally, we will discuss the implications of our findings for the collaborative and automatic entrainment theories, and highlight

---

[1] A hypernym is the superordinate concept of another word, for example, animal is hypernym of cow, and organ is hypernym of lung and stomach.

[2] Hyponymy is the opposite of hypernymy: lung and stomach are both hyponyms of organ.

some issues for further research.

2 Reference in Semantic Theory

The choice of words interlocutors employ to refer to a given object is greatly influenced by extra-linguistic contexts. Our view on reference follows cognitive accounts of semantics (Langacker, 1987; Feldman, 2006), which define a speaker's embodied knowledge of the world and resulting cognition as the foundation of semantics. This view is opposed to the view that the choice of words in an utterance is a function of syntactic or semantic selectional restrictions, in which a word limits the words that can accompany it (Chomsky, 1969). A basis for much research in cognitive semantics is Olson's (1970) statement that "everything has many names and every name 'has' many things." (p. 162). Thus, the relation between words and referents is not a direct relation but is mediated by the context. The mediating component determining the function of a word can be the experience of perceiving objects in a context. It is, therefore, not possible to define 'the meaning' of a word that holds for all contexts, but "the meaning of a word is its use in the language" (Wittgenstein, 1958). In this sense, the main factor influencing the choice of reference "is made so as to differentiate an intended referent from some perceived or inferred set of alternatives" (Olson 1970).

This view is also supported by the contrast set model (Dale and Reiter, 1995), where the choice of words in a referring expression is made in order to rule out the other possible referents within the given physical context.

Besides the context, the addressee plays an important role for the choice of words: A Speaker distinguishes between information she considers given, i.e. information she thinks the listener should already know and accepts as true, and information the speaker considers new, i.e. which she thinks the listener does not yet know. But speakers not only take into account what they think the listener knows, they also expect the listener to make inferences from shared knowledge, which is called common ground (Clark & Bangerter, 2004). This can either be information that is publicly known or joint personal experiences, e.g. items that are perceptually co-present. Nevertheless, "common ground isn't a homogeneous body of well-established propositions" (Clark & Bangerter 2004, p. 35), it is rather changing all the time in the course of a conversation and far from being totally clear to both interlocutors, since it is uncertain whether some propositions belong to common ground or not. So, a conversation can be seen as establishing and testing out common ground all the time, which only works if both interlocutors work together. Clark and Wilkes-Gibbs (1986) studied the collaborative nature of referring in an experiment where participants had to work together in a referential communication task, one as director and one as matcher. During six trials the director had to get the matcher to arrange twelve cards showing Tangram figures in a specific order. Clark and Wilkes-Gibbs (1986) found that participants needed fewer words and fewer turns per figure in the course from trial one o six. As an explanation they argue that director and matcher collaborate with each other to develop an appropriate reference system. According to Clark and Bangerter (2004) interlocutors initiate a process that has two goals:

- Identification – Speakers want their addressees to identify a particular figure under a particular description.
- Grounding – Both interlocutors try to establish the mutual belief that the addressee has identified the referent well enough for current purposes.

Grounding was defined by Clark and Bangerter (2004) as establishing a thing "as part of common ground well enough for current purposes". An important question that now remains is how the development of common ground actually works. This will be discussed in the following section on *lexical entrainment*.

3 Lexical Entrainment

One process that can be seen as a part of the collaborative behavior that interlocutors show in a dialogue is lexical entrainment (LE). In the course of this linguistic adaption speaker and hearer converge on shared terms. A sample definition is expressed in the following:

> "[I]f *A* talks to *B* and uses a term such as *pointer* to refer to an [sic!] graphically displayed object, i.e. leads in the usage of the term – and *B* (from then on) also employs the term, i.e. follows lead of *A*, then we have a classic case of entrainment." (Porzel, 2006, p. 1)

Every time a speaker selects words to refer to an object, he or she assumes a conceptual perspective for the listener to adopt with regards to the given referent. If there is need to refer to the same object again, interlocutors will generally make use of the same referential conceptualization by reusing the same term(s) or an abbreviated version (Van der Wege, 2009).

Two distinct views of lexical entrainment have emerged: The mechanistic model sees LE more as an automatic process, while the collaborative view emphasizes on strategic cooperative aspects of it. According to the mechanistic model of LE (Pickering & Garrod, 2004), the linguistic representations used to understand and to produce utterances by two interlocutors become automatically aligned on several levels, not just in the syntactic, lexical and phonological elements, but even on the situation model in discussion. Alignment is supposed to "work via a priming mechanism, whereby encountering an utterance that activates a particular representation makes it more likely that the person will subsequently produce an utterance that uses that representation" (Pickering & Garrod, 2004).

Conversely, for proponents of the collaborative view (Clark & Wilkes-Gibbs, 1986), lexical entrainment is part of the conscious, collaborative effort to achieve *identification* and *grounding* (see above). An important requirement for making a reference that the interlocutor will understand is the establishment of mutual knowledge. The first person who makes a reference has to be convinced that the identity of the referent is truly going to become part of the common ground of both interlocutors. The second speaker, trying to understand the reference, should let the first one know, whether or not he/she understands it. One way of achieving this is by using the same expression in the further course of the dialogue. Hence, lexical entrainment can rather be understood as a conscious or strategic process. We will now have a look at the factors that influence this process.

3.1 Factors that Influence Entrainment

Following the collaborative approach, lexical entrainment is regarded to be based on two principles, the Principle of Contrast and the Principle of Conventionality (Clark, 1988; Van der Wege, 2009). These are also the primary principles children employ when learning new words.

According to the Principle of Contrast, children act on the assumption that any difference in form of a word indicates that there is a difference in meaning. The Principle of Conventionality says that for certain meanings a conventional form exists. When one does not use this form that speakers of a community expect to be used, there has to be a reason, like having another, contrasting meaning in mind.

Van der Wege (2009) applies these principles for the field of reference in general. The principles can be applied to the language of a community, as well as to one single conversation. New words are seen in contrast with words that are already known or have already been established in the course of the conversation. Van der Wege (2009) assumes that not only word meanings are contrasted by speakers but also the words they use in their referring expressions and the conceptualizations of the referent that underlie their choices.

By using this term, she intends to leave open that the linguistic precedents used and maintained by the speakers might be conceptual, rather than linguistic. Following these principles, we can firstly predict a strong preference of speakers to continue using an established conceptualization when referring to the same referent. For example, a speaker who started to refer to a particular shoe as a 'black loafer' will continue to call it 'black loafer' when referring to it again instead of choosing a new reference phrase like 'shoe' (Brennan & Clark, 1996). Secondly, Van Der Wege (2009) predicts lexical differentiation: When referring to a new referent, there should be a "strong preference to use a reference phrase and corresponding referential conceptualization that contrasts with other previously established referential conceptualizations." (p. 449)

Another factor that influences the choice of a referring expression is the context of established references within a conversation. When referring to the same referents multiple times with the same conversational partner, speakers often underspecify referents, as in the following example of Clark and Wilkes-Gibbs' (1986, p. 12):

1. a person who's ice skating, except they're sticking two arms out in front
2. the person ice skating that has two arms
3. the person ice skating, with two arms
4. the ice skater

Even if the shortened references (such as ’ice skater’) would be ambiguous for a third person, in the historical context of the conversation the reference will be clear for both interlocutors.

When the set of potential referents changes, so that an established conceptualization is no longer unambiguous, speakers may be uncertain between maintaining the established conceptualization and trying to be as unambiguous as possible in the current context (Van der Wege, 2009). This means that the speaker may consciously consider changing the choice of words to adapt to the new context.

In the present study, we wanted to find out what kind of influence semantic distance in form of different hypernyms or hyponyms has on the process of lexical entrainment. A further question was if the results would speak rather in favor of the mechanistic or the collaborative account on LE.

4 Materials and Methods

First, we will discuss our definition of semantic distance which starts from the concept of hypernymy/hyponymy. Furthermore, the experimental setup and execution are discussed in the subsequent section. The last section deals with the analysis of the data and the program that was implemented for this purpose.

4.1 Hypernymy and Semantic Distance

Hypernymy and Hyponymy are two different ways in which word senses can be related. For example, *animal* would be the hypernym of *bear* and *walrus*, conversely they are the hyponyms of *animal*. A word can also be a synonym to its hypernym in contexts where it is used to specify the same intended referent: ”Thus, ’I took your money’ is synonymous with ’I took the five dollars’ if the five dollars is your money.” (Olson, 1970, p. 267).

We based our concept of semantic distance on hyponymy and hypernymy relationships in the lexical database of English WordNet 2.1 (Budanitsky & Hirst, 2001; Gurevych & Niederlich, 2005). We started by identifying how many common hypernyms two nouns have in WordNet 2.1. This semantic distance is a measure to compare whether two nouns are more or less similar than two other nouns. As an example we take the following three pairs of nouns:

Fish - Fish

Whale - Giraffe

Dinosaur - Butterfly

”Fish” obviously has the same amount of hyponyms ”fish” has, so this is the most similar two nouns can get, i.e. identical (Table 1). The semantic similarity of ”whale” and ”giraffe” is given by the first hypernym in which they coincide, which is, according to WordNet, ”placental”. ”Placental” has 9 hyponyms so ”whale” and ”giraffe” would have a measure of 10 in the practical way of comparing them with the semantic similarity of ”dinosaur” and ”butterfly”. ”Dinosaur” and ”butterfly’s” first common hypernym is ”animal” that has 5 hyponyms. Following from the amount of common hypernyms we developed 3 levels of semantic similarity: Level 1 included words which have all hyponyms in common in addition to the same definition, which meant that both participants had the same image on their card. For Level 2, the definition in WordNet had to be different and most of the terms had at least one uncommon hypernym. The terms for Level 3 had at least two uncommon hypernyms. The final categorization was performed by four raters in a separate evaluation experiment. All raters categorized perceived semantic distance of pairs of images. Only those card sets were included where all raters agreed on the semantic distance.

Table 1: Levels of semantic similarity and examples.

Semantic Similarity	**Noun 1**	**Noun 2**	**Common Hypernyms**
1	Fish	Fish	Absolute similarity = identity
2	Whale	Giraffe	Placental, Mammal, Vertebrate, Chordate, Animal ...
3	Dinosaur	Butterfly	Animal, Organism, Living thing, Physical Object ...

4.1.1 Setup and Execution

The general experimental set-up was inspired by Clark & Wilkes-Gibbs (1986), but the content and the number of the cards were different. The original task was restricted to ordering identical sets of cards in the same order. In the present experiment, the images on the corresponding cards were semantically related, but not identical, making reference harder and enforcing the development of a matching strategy.

Two participants were seated opposite to each other. They were separated by a wall, so that they could not see each other's cards. Each one got five cards in a sequence randomly chosen by the experimenter; whereby for each card of one participant, there was one semantically related card in the partner's set. They were told that the goal was to arrange their cards in the same sequence as the other participant. One of the participants was the 'leader', which means he had to lead the other one ('follower') to arrange the cards in the correct sequence. This role was alternated after each completed sequence. Each couple did four trials in arranging the cards in one of the three levels. Throughout the four trials, the same cards were used.

In total, 50 people (25 couples, randomly paired) were asked to take part in the experiment. They were all students between 18 and 31 years old. The experiment was introduced to the participants as a "card game" and "only a warm-up" for another subsequent experiment to make sure that the participants would not care too much about what they said.

The speech of the participants was recorded and transcribed. 18 text files were considered useful data as input for the analysis, as some of the data had to be rejected due to some participant's lacking knowledge of German or other complications. Four of these 18 couples had done Level 1 of semantic distance; seven, respectively, had done Level 2 and Level 3.

4.2 Analysis

A program was implemented to process the transcription of the recorded conversations. First, two kinds of results were analyzed:

- Amount of words used by the two participants for every trial.
- Frequency of all the nouns uttered which referred to the content depicted on the cards, distinguishing whether they were hyponyms or hypernyms.

Moreover, the course of entrainment during the 4 trials was analyzed. Therefore it was counted as entrained reference, when a speaker used a referring expression that had been used before by the other speaker; in which usage by the other speaker may have been at any prior point in the experiment. We also counted a reference as entrained when it could unambiguously be identified as a reduced form of a reference the other speaker had used before (e.g. "the soccer goal" = "the goal"). A referring expression used by a speaker was considered non-entrained when it had not been used before by the other speaker in the whole experiment. When a new referent was introduced, the reference was always non-entrained.

5 Results

The following graph shows how the average number of words uttered by each of the participants varied on each of the different levels and trials. Generally, the further the semantic distance between the cards, the more words participants needed in order to solve the task, as shown in Figure 3. For Level 3 the number of words was much higher for the first two turns compared to Level 1 and Level 2. For Level 2, there were still a lot more words needed than for Level 1. But, in the course of the interaction, the dispersion of the number of words on the three levels decreased.

While for turn one, the average amount of words in Level 2 is almost 50 times higher (990 words) compared to Level 1 (19 words) and 134 times higher (2015 words) in Level 3, for Trial 4 the amount of words is below 250 for all levels with relatively small differences between the different levels.

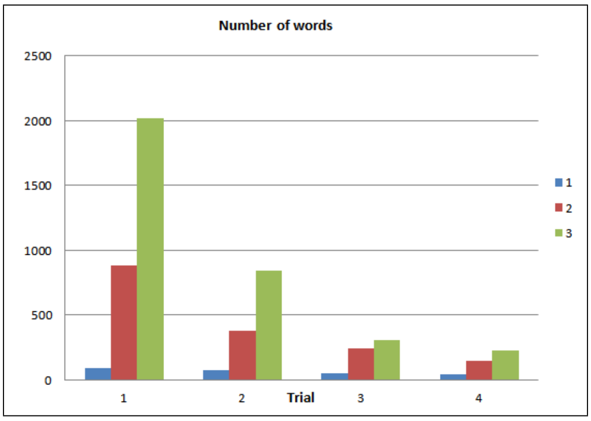


Figure 1: Average number of words in the course of the four trials for Level 1 (blue), 2 (red) and 3 (green).

Figure 2 shows the percentage of hypernyms for all cases where referring expressions were used, i.e. all cases of words with which the participants referred to the objects on the cards.

While for Level 1 the rate of hypernyms is 0% throughout all trials – which means that, as was ex-

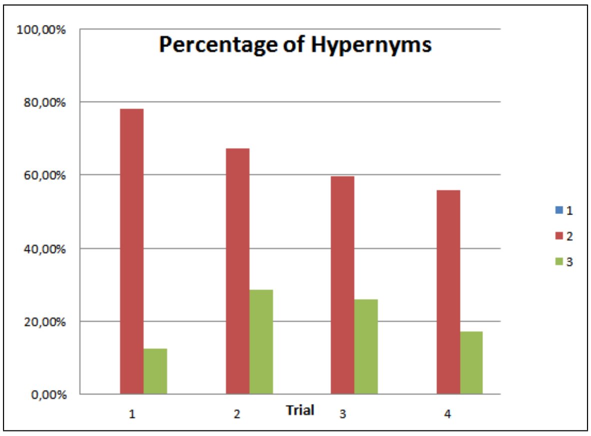


Figure 2: Percentage of Hypernyms per Trial in Level 2 (red) and Level 3 (green).

pected, no hypernyms at all were used, for Level 3 the number of hypernyms increased in the second trial and then decreased in the third and fourth trial. For Level 2 a slight overall decrease from Trial 1 to Trial 4 can be observed.

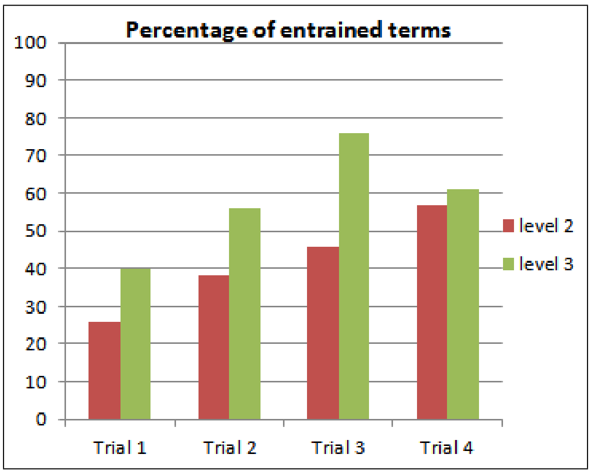


Figure 3: Percentage of entrained terms (of all referring expressions) in Level 2 (red) and Level 3 (green).

The analysis of the number of entrained terms shows that for level 2 slightly more entrained terms are used in each trial (from 26% in trial 1 to 57% in trial 4) while in level 3 the percentage of entrained terms first increases and then drops from 76% in trial 3 to 61% in trial 4 (s. figure 3).

We will now provide some samples from our data. Participants that show a low percentage of entrainment did not entrain for the majority of the terms used. Example 1 gives evidence that entrainment does not occur in all cases (i.e. not for all referring expressions used by a participant pair). Example 1 shows all utterances referring to the pair SHIRT – PANTS in the dialogue between Participants A and B, who take turns in taking the role of leader (L) and follower (F). The example is from Level 2, having an intermediary semantic distance between corresponding terms. Certain references are not entrained at all: each participant is stating what is on their card, and relying completely on the partner to perform the matching. We call this *lexical non-entrainment*.

Example 1

Trial 1
A (L): Ich hab ganz links, äh, ’n **Hemd**. Ich hab ganz links, ja... Äh, dann kommt ’ne Ananas als nächstes.
*I have on the very left, [HES], a **shirt**. I have completely left, yes, [HES] then comes a pineapple next.*

Trial 2
B (L): Dann hab ich, äh, Orangen. Ja, Orangen.
Dann hab ich Delfin. Und ’ne **Hose**.
Then I have [HES] oranges, yes, oranges.
*Then I have a dolphin and some **pants**.*

Trial 3
A (L): Also, ganz links hab ich wieder das **Hemd**.
*Well, on the very left I have the **shirt** again.*

Trial 4
B (L): Ich hab jetzt Delfin, und dann den Baum. **Hose**, ähm, Baseball und Orange.
*Now I have dolphin and then a tree. **Pants** [Hes] baseball and orange.*

This effect did not occur in the Level 1 data, as the initial referential expressions used were already identical.

Example 2 shows all utterances relating to the pair FOOTBALL GOAL – BASEBALL BAT from Level 3 between participant A and B who take turns in taking the role of leader (L) and follower (F):

Example 2

Trial 1

A (L): Äh, rechts Walnuss, Schmetterling, **Tor**, Schuh und, äh, – Gott was ist das? – irgend'ne Pflanze.
*[Hes], on the right walnut, butterfly, **goal**, shoe and [Hes] – God what's that? – some kind of plant.*

B (F): Ah, ok. Sag nochmal, Chef.
Ah, ok. Say it again.

A (L): Äh, Walnuss, Schmetterling oder Motte, eins von beiden. **Fussballtor, Handballtor, irgend'n Tor**, ein wunderschöner Schuh.
*[Hes], walnut, butterfly or moth, one of both. **Football goal, handball goal, some kind of goal**, a lovely shoe.*

Trial 2

B (L): **Baseballkeule.**
Also, was anzuziehen, 'ne Pflanze, ne? Frucht, Tier, **Sportgerät**.
Baseball bat.
*So, something to dress, a plant, right? Fruit, animal, **sports equipment**.*

Trial 3

A (L): Also, ganz vorne **das Sportgerät, Tor**, Pflanze, äh, fleischfressende Pflanze. Ähm, die Walnuss, der Schuh und der Schmetterling.
*So, right ahead the **sports equipment, goal**, plant, [Hes], carnivorous plant. [Hes], the walnut, the shoe and the butterfly.*

Trial 4

B (L): Dinosaurier, Gurke, Baum, **Baseballschläger**, Mütze.
*Dinosaur, cucumber, tree, **baseball bat**, cap.*

As can be seen in this example, from Trial 1 to Trial 4 fewer words are needed to reach the goal in each turn. In Trial 1, the participants have not reached the stage of entrainment yet, the hyponym 'goal'/'football goal' is used. In Trial 2, participant B introduces the hypernym, 'sports equipment', to make sure that each of them has understood what the task is about. Participant A follows the usage of the hypernym in Trial 3. Nevertheless, in Trial 4 B goes back to the hyponym. The hypernym is not necessary for their communication, as both know that they are indirectly referring to it. This phenomenon we term *lexical de-entrainment.*

It is clear this process cannot happen in that data obtained in our Level 1 experiments, as the entrained terms were identical to the initial referential expressions used. Thus, there was no need to go to another lexical expression in the first place, and therefore no way to return to an initial state. In Level 3 data, we observed this process in 4 out of 7 dialogues, i.e. the percentage of entrained terms decreased from trial 3 to trial 4. Example 3, taken from level 3, illustrates *lexical de-entrainment* again.

Example 3

Trial 1

[...]

A(L): dann habe ich ... eine **Pflanze**
*I have .. a **plant**.*

B(F): mm, ja, ich habe einen **Baum**, vielleicht ist das so ein bisschen das gleiche.. und ich habe einen **Hut**, das ist vielleicht...
*[HES] yeah I have a **tree**, maybe that's kind of the same, and I have a **hat**, maybe that's...*

A(L): Das ist schon mal gut, dann habe ich ein **Sportgerät** als **Tor**.
*That's very good already; I also have **sports equipment** as **goal**.*

[...]

A(L): Das ist schon mal gut...und **Schmetterling** als **Tier** vielleicht.
*That's already good, and a **butterfly** as **animal** maybe.*

[...]

Trial 2

B(L): Also, erst das **Tier**.
*So, first the **animal**.*

A(F): Ja.
Yeah.

B(L): Dann die **Pflanze**, **Kleidungstück**.
*Then the **plant**, **clothing piece**.*

[...]

B(L): Das **Sportsding** und dann das **Essens-** ... mh
*The **sports thing** and then the **eating** [HES]*

Trial 3
A(L): Am Anfang habe ich den **Schmetterling**, das **Tier**.
*At first I have the **butterfly**, the **animal**.*

B(F): Ja.
Yeah.

A(L): Das **Sportsgerät**, das **Tor**.
*The **sports equipment**, the **goal**.*
[...]

A(L): **Kleidungsstück**, den **Schuh**.
***Clothing piece**, the **shoe**.*
[...]

A(L): Dann die **Walnuss** als **Nahrung**.
*Then the **walnut** as **food**.*
[...]

A(L): und die **Pflanze**.
*And the **plant**.*

Trial 4
B(L): Ok **Mütze**, **Baum**, **Dinosaurier**, **Gurke** und **Baseballschläger**.
*Ok **hat**, **tree**, **dinosaur**, **cucumber** and **baseball bat**.*

Participant B did not address the entrained hypernyms in Trial 4 to accomplish the goal. She referred to the cards she held in her hand ignoring the already entrained hypernyms and knowing that her cards were different to A's. She *de-entrained.*

6 Discussion

In our experiments we found – as could be expected – that the larger the semantic distance between corresponding cards, the more words per trial were needed to perform the task. Also, the decrease in word number from Trial 1 to Trial 4 was higher for the Levels 2 and 3 than within Level 1.

This directly reflects the higher collaborative effort needed to establish common ground. While for semantically closely related objects the listener could easily infer which object the speaker was referring to, with a high semantic distance the knowledge of which objects correspond to each other needed to be built up during the task. By the third or fourth trial, in most cases common ground had been fully established, therefore the dispersion of the number of words on the three levels decreased.

At this point both participants usually knew to which objects they were referring and the words uttered did not matter in order to complete the indicated task. In this way, the expressive distinction between hypernym and hyponym had been overcome. In some cases, such as Example 1, participants reached this stage very soon and therefore did not need to rely on lexical entrainment at all for solving the task. In other cases, after entrainment had been used for establishing common ground, at some point lexical de-entrainment occurred, as the previously entrained terms were not needed anymore.

In an analogy to a Hegelian Spiral one can think of de-entrainment as starting with an initial state of non-entrained terms, which – through processes of alignment – turns into a second state in which terms become more and more entrained. Lastly, speakers can reach a third state where terms become de-entrained again, which looks on the surface almost identical to the first level, but now a crucial conceptual change has occurred in the interlocutors' understanding of the de-entrained terms. In some cases, however, the step of lexical entrainment can be skipped, reaching the third stage of common ground directly.

The theoretical implications of these findings are clear: If entrainment was to be an automatic process based on basic priming and joint action principles there would be no reason or even mechanism to trigger de-entrainment processes. This means that speakers would remain in phase two which would reinforce itself more and more through automatic processes. If we are dealing with collaborative strategies that serve multiple goals, e.g. mutual understanding as well as economy, scenarios can be envisioned in which it becomes feasible to drop previously entrained terms for the sake of one's own cognitive economy without putting mutual understanding at risk. Or, in the more extreme case, mutual understanding may be reached so early in the dialogue that the process of entrainment is simply not necessary. In our opinion, such a scenario is manifested in our data.

When it becomes conceptually evident that each speaker has only one instance of the hypernym at hand, e.g. A has an *apple* and B has an *orange* as instances of the hypernym *fruit* – then A's apple becomes the conceptual counterpart of B's orange. As a consequence, the previously established and entrained hypernym *fruit* can be abandoned, because A knows that B will understand his reference to the apple to refer to her orange.

In the data set we observed a decline in the number of hypernyms used after a while, which is – in our minds – insurmountable with an automatic view on entrainment, since this view would predict at rising or at least a constant level of entrained terms.

6.1 Future work

The phenomenon of lexical de-entrainment should be studied further with larger-scale studies. In order to quantify the de-entrainment level a conservative metric could take the maximal level of entrainment and calculate the integral between the actual decline of the curve and an assumed constancy at that level. This integral, therefore, quantifies the level of de-entrainment over time, based on the prior level of entrainment. Having more data would also enable to give comparative metrics concerning the slopes of the entrainment and de-entrainment curves.

A further goal should be to gain further insight into the specific conditions that cause lexical de-entrainment in order to get a better understanding of the relationship between the collaborative striving for mutual understanding, and the desire to save cognitive effort.

7 Acknowledgements

Funding by the DFG for the SFB/TR 8, project I5-[DiaSpace], is gratefully acknowledged. Further we would like to thank the students who participated in our study.

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), 1482-1493.

Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA* (pp. 29-34).

Chomsky, N. (1969). *Aspects of the theory of syntax.* Cambridge, Mass.: M.I.T. Press.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, 15, 317-335.

Clark, H.H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Clark, Herbert H. & Bangerter, A. (2004). Changing ideas about reference. In D. Sperber & I.A. Noveck (Eds.), *Experimental Pragmatics*. Hampshire, New York: Palgrave Macmillan.

Feldman, J. (2006). *From molecule to metaphor: A neural theory of language.* Cambridge, MA: MIT Press.

Fellbaum, C. (1994). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Furnas, G., Landauer, T., & Dumais, G. (1987). The vocabulary problem in human-system-communication: an analysis and a solution. *Communications of the ACM*, 30 (11), 964-971.

Gurevych, I. & Niederlich, H. (2005). Accessing GermaNet data and computing semantic relatedness. *Companion Volume of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 5-8).

Herrmann, T. & Deutsch, W. (1976). *Psychologie der Objektbenennung.* Bern: Hans Huber.

Herrmann, T. & Grabowski, J. (1994). *Sprechen: Psychologie der Sprachproduktion.* Heidelberg: Spektrum Akademischer Verlag.

Langacker, R. (1987). *Foundations of cognitive grammar.* Stanford: Stanford University Press.

Olson, David R. (1970). Language and thought: aspects of a cognitive theory of semantics. *Psychological Review*, 77, 257-273.

Pickering, M. J. & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169-226.

Porzel, R., Scheffler, A. & Malaka, R. (2006). How entrainment increases dialogical effectiveness. *Proceedings of Workshop on Effective Multimodal Dialogue Interfaces*, Sydney.

Van der Wege, M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60, 448-463.

Wittgenstein, L. (1958). *Philosophical investigations.* Oxford: Basil, Blackwell & Mott.

Dale, R., & Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 19(2), 233-263.

The Earliest Utterances in Dialogue: Towards a Formal Theory of Parent/Child Talk in Interaction*

Jonathan Ginzburg and Sara Moradlou
CLILLAC-ARP (EA 3967) & Laboratoire Linguistique Formelle (LLF) (UMR 7110)
& Laboratoire d'Excellence (LabEx)–Empirical Foundations of Linguistics (EFL)
Université Paris-Diderot, Sorbonne Paris Cité, Paris, France
`yonatan.ginzburg@univ-paris-diderot.fr, sara.moradlou@gmail.com`

Abstract

Early, initial utterances by children have received relatively little attention from researchers on language acquisition and almost no attempts to describe them using a formal grammar. In this paper we develop a taxonomy for such utterances, inspired by a study of the Providence corpus from CHILDES and driven by the need to describe how the contents of early child utterances arise from an interaction of form and dialogical context. The results of our corpus study demonstrate that even at this early stage quite intricate semantic mechanisms are in play, including non-referential meaning, akin to non–specific readings of quantifiers. We sketch a formal framework for describing the dialogue context and grammar that underlies such utterances. We consider very briefly and informally how some such utterances emerge from parent/child interaction.

1 Introduction

The early stages of a process are crucial in understanding its subsequent development. Although there has been some work in this area, which we summarize below, it seems true to say that the early, initial utterances by children have received relatively little attention from researchers on language acquisition and almost no attempts to describe them using a formal grammar.[1] Given that parents and carers can make sense of much of what young children say, we assume the mechanisms of this understanding process deserve formal analysis and, unless compelling reasons to the contrary be given, incorporation within some notion of grammar. It is clear that such a notion will rely, even more than is the case for adult spoken interaction, on a detailed theory of context.

In this paper we develop a taxonomy for early child utterances. In contrast to previous work, summarized in section 2 which was strongly based on speech act theory and paid little attention to the fine structure of semantic combinatory mechanisms, our own taxonomy, developed in section 3, based on the Providence corpus from CHILDES (Demuth et al., 2006), is driven by the need to describe how the contents of early child utterances arise from an interaction of form and dialogical context. The results of our corpus study, described in section 4, demonstrate that even at this early stage quite intricate semantic mechanisms are in play, including non-referential meaning, akin to non–specific readings of quantifiers. In section 5 we sketch a formal framework for describing the dialogue context and grammar that underlies such utterances, showing that even at this initial stage, the child grammar is in a sense continuous with adult grammar. In section 6, we consider very briefly and informally how some such utterances emerge from parent/child interaction.

2 Literature Review

Previous work on categorizing children's utterances is mainly based on Speech Act Theory (Searle, 1969; Austin, 1975). The work on speech act analysis of child language attempts to characterize the nature of parent-child interactions and its links to language learning. These approaches to language acquisition view verbal forms as means

---

*We acknowledge support from Lab(oratory of)Ex(cellence)–EFL (ANR/CGI), in particular for providing a graduate fellowship to the alphabetically second author. We would like to thank Eve Clark, Judit Gervain, and three anonymous reviewers for DialDam for very useful comments on an earlier draft.

[1] For brief discussion in the context of a proposal concerning the evolution of grammar, see (Jackendoff and Wittenberg, 2014).

of expressing communicative intents and emphasize the role of function (e.g., (Bates, 1976; Bloom, 1967; Bruner, 1975; Dore, 1975; Dore, 1974; Ninio, 1992; Ninio and Bruner, 1978; Ninio and Snow, 1988)) as opposed to viewing learning as a mapping from form to meaning. Thus, there is no attempt to describe the fine structure of descriptive contents, which as we will see, already involves quite intricate combinatorial mechanisms. Nor is there explication of how these arise drawing on contextual information.

Ninio (1992) shows a strong correlation between single word utterances of children and those of their mothers, indicating a high degree of form-function specificity (see also, (Bruner, 1975; Shatz, 1979; Snow, 1972)). This co-variation between parent speech and child's utterances is the basis of most acquisition theories that consider conversation as an important factor. Bruner (1981) notes that mothers used highlighting or fore-fronting of the objects extensively, when introducing them to children, and also routinized individualized characteristic ways of preparing for presentation when the child was not attending to them (e.g., calling by name). Based on his observations that the interactions between mother and child follow highly regular patterns and that these patterns evolve as the child becomes more and more competent in language use, Bruner (1981) argues that this *Language Assistance System* plays an important role in children's language development.

We briefly describe here two classification schemata based on the speech acts approach we mentioned above: Dore (1974) sets out to explain the development of adult speech acts repertoire; using the data collected in a longitudinal study of two subjects in their single-word stage of language production, he categorizes *Primitive Speech Acts* of these infants into 9 types each of which differs with the others at least in one feature, either in form or function. INCA[2] (Ninio and Wheeler, 1986) and its abridged version (INCA-A) (Ninio et al., 1994) are annotation schemas that code communicative intents in two parts: level of verbal interchange, which is defined as a series of utterances that serve a unitary interactive function, (e.g., negotiating immediate activity, discussing joint focus of attention, etc) and utterance level speech acts (e.g., requesting, proposing an action, etc.). These systems are meant to code communication attempts for both adults and children in different stages of acquisition; this makes the number of types to chose from quite big for the coder and the annotation work rather difficult.

Next, we describe our taxonomy for classifying children's early utterances, with annotation effort, and dialogue dynamics in mind.

3 Corpus Study

We annotated the odd number files from 11 to 15 months for Naima and Lily of Providence corpus (Demuth et al., 2006) in CHILDES (MacWhinney, 2000) with utterance types based on Fernández and Ginzburg (2002)'s taxonomy of adult non-sentential utterances (NSU).[3] However, these adult NSU types do not cover all of the NSUs observed in child language; furthermore, some adult NSU categories do not occur at all in the early stages of acquisition. We developed our taxonomy for the early stage of child language using Naima's utterances in her one-word stage: we manually categorized the utterances into one of the types we will discuss shortly, based on their form and the conversation function they served, trying to maximize the number of phenomena covered by our taxonomy. We only retained the types that occurred in more than 4 percent of the utterances in at least one of the observation sessions. The motivation for this was to exclude utterances that occurred very rarely; we applied the threshold frequency on sessions instead of the complete development set to capture development of types over time.

Below we describe each class of utterances with examples; these classes are organized in three broad categories: labeling types are the utterance types that refer to the visual scene. The second class of types are those that follow up on parent's utterances. Attention directing types are initiated by the child and play a role in managing joint attention.

3.1 Labeling types

Visual Object Pointing

We classified word or word-like utterances that referred to entities in the visual field in order to label them as **VisObjP**. This type was usually accompa-

---

[2] Inventory of Communicative Acts

[3] We used the Providence corpus since it is multimodal and children are recorded there from very early stages of speaking.

nied by pointing (example 1) or reaching-out gestures (example 3), and gaze towards the object that was being named.

(1) CHI: yyy .
pho: ˈkʌkə
sit: CHI is pointing to the microphone
MOT: yeah , that's my microphone .
MOT: you have one too . (Naima at 0;11.28)

(2) CHI: yogurt .
MOT: yogurt .
MOT: is it good ? (Naima at 1;0.28)

(3) CHI: bike yyy .
pho: ˈbæk ˈbæbæ
sit: CHI picks up toy bike
MOT: oh your bike . (Naima at 1;1.25)

Visual Predication

When a property of an entity or an event in the visual situation was predicated by an utterance, it was annotated with **VisPred**. This type is similar to VisObjP in the co-occurring gestural-proxemic cues, and in its dependence on the visual situation.

(4) MOT: hey , you dropped an odio . CHI: down . sit: CHI is picking up the cereal. (Naima at 0;11.28)

(5) sit: book reading activity. CHI: baby . MOT: and there's a baby . CHI: sleeping . pho: ʃLiːpi MOT: yes the baby is sleeping . MOT: on the bed . (Naima at 1;2.23)

(6) CHI: big . pho: ˈbiːgə sit: CHI is trying to put a toy chicken into a cup. MOT: oh the chick is too big for that small cup . MOT: it doesn't fit in . (Naima at 1;3.12)

Visual Onomatopoeic Utterances

VisOno class covers onomatopoeic utterances (animal, machine, and though less often, human sounds like imitations of laughing and crying) when it is triggered by an entity or event in the visual attention of the child:

(7) CHI: baa baa baa baa baa baa baa . sit: CHI is pointing at a sheep in a picture book. MOT: that's right , that's the sheep going baː baː . (Naima at 1;0.28)

3.2 Types following up on parent's utterances

Short Answer

Utterances in response to parent's Wh-questions or implicit Wh-questions (usually using pausing, intonations, and gestural cues) were categorized as **ShortAns**. This was irrespective of the correctness of child's answer, as long as the utterance was word-like and it could be interpreted as an answer taking into account child's non-verbal behavior and parent's interpretation of the utterance:

(8) MOT: who's that coming in the door? CHI: Daddy . MOT: yes that's right . (Naima at 0;11.28)

(9) MOT: what else is here ? CHI: duckling . pho: ˈgʌkəlɪ MOT: duckaling , that's a duck . MOT: quack quack , (.) and + . . . (Naima at 1;0.28)

Repetition Acknowledgment

This class contains utterances that follow up on parent's previous utterance(s) by repeating (part of) it. **RepAckWord** was distinguished from pure imitations (**Imit**) based on child's participation in the conversation and annotators' judgment of the nature of the repetition: utterances that functioned purely as practice for pronunciation were annotated as **Imit**. In addition, if child repeated an utterance that was not directed to her or she did not seem to be paying attention, that utterance was not considered a repetition acknowledgment but as an imitation (compare 12 and 10). We also considered onomatopoeic utterances that were related to parents previous utterance (for example, imitating an animal that has been mentioned by the parent) as repetition acknowledgment and tagged them with **RepAckOno**.

(10) MOT: that baby has a bottle , did you notice that ? CHI: bottle . MOT: yeah , baby has a bottle . (Naima at 1;0.28)

(11) MOT: it's a shovel . CHI: shovel . MOT: just like your shovel . (Naima at 1;3.12)

Imitation

(12) MOT: I went to +//. FAT: xxx . MOT: pain d'avignon +//. CHI: yyy . pho: ˈpliː MOT: pain d'avignon yesterday . CHI: yyy . pho: ˈpliː MOT: play . FAT: wash xxx first . MOT: brioche bread and some +//. CHI: brioche yyy yyy . (Naima at 1;2.23)

3.3 Attention directing types

Call

This category contains instances of "Daddy" and variations of "Mommy" used as a means of directing or establishing shared attention:

(13) CHI: Mama ? MOT: yeah ? MOT: that's my tea , you can't drink my tea , babies don't drink tea . (Naima at 1;0.28)

Request

The utterances that functioned as requests for entities using surface forms analogous to those in VisObjP, as in (14), were annotated with **ReqObj**, whereas the requests with forms referring to events or properties of entities were tagged as **ReqPred**, like in (16) and (15).

(14) CHI: Mommy . MOT: yes Naima . CHI: water . MOT: you want some more water . (Naima at 1;1.25)

(15) CHI: more . pho: ˈmːon MOT: oh more ? MOT: okay , here's a big piece of wheat . MOT: put in that one . MOT: make sure you chew that , okay ? (Naima at 1;0.28)

(16) CHI: Daddy . FAT: yes baby . FAT: you look so serious and earnest . CHI: up . CHI: Daddy . MOT: up . FAT: up oh . FAT: okay baby . FAT: you said up . FAT: pick me up ? (Naima at 1;2.23)

Types		Ages				
		0;11.28	1;0.28	1;1.25	1;2.23	1;3.12
Vis	ObjP	43.3	30.5	32	32.3	25.6
	Pred	7.8	0.9	2.7	2.4	8.1
	Ono	0	1.7	1.3	5.7	1.9
RepAck	Word	3.3	22.9	14.7	13.7	13.4
	Ono	1.1	5.1	0	2.4	3.1
Imit		10	4.2	6.7	4	4.7
ShortAns		10	12.7	12	4.8	11.6
Call		11.1	6.8	18.7	17.7	7.2
Req	Obj	0	0.9	4	9.7	2.5
	Pred	0	1.7	0	1.6	4.7
Multi-word		0	0	0	1.6	7.5
% covered		86.7	87.3	92	96	90.3
# annotated		90	118	75	124	320
# unintelligible		162	104	66	142	193
# total		299	245	171	356	636

Distribution of utterance types over time, in percentage.
Ages in years;months.days

Table 1: Distribution of types for Naima

Types		Ages				
		1;1.02	1;2.02	1;2.30	1;3.27	1;4.25
Vis	ObjP	10.3	18.2	0	0	24.1
RepAck	Word	24.1	9.1	0	0	6.9
Imit		13.8	4.5	10	0	6.9
ShortAns		48.3	59.1	90	71.4	55.2
Call		0	4.5	0	28.6	6.9
% Covered		96.6	95.5	100	100	100
# annotated		29	22	10	7	29
# unintelligible		107	215	468	723	345
# total		148	234	478	732	378

Distribution of utterance types over time, in percentage.
Ages in years;months.days

Table 2: Distribution of types for Lily

4 Results

Table 1 summarizes the type frequencies in Naima's speech for the types retained in the taxonomy after applying the above mentioned threshold to remove very rare types. In this table, percent covered is the ratio of utterances that our taxonomy could account for, over total number of annotated utterances. It is worth noting that we only annotated the first instance of an utterance when it was repeated multiply in a sequential manner. We also applied this metric to a new set of transcripts, odd files of Lily from the Providence corpus: our taxonomy achieved high coverage for both Naima (86.7–96%) and Lily (95.5–100%). Distribution of types for Lily is provided in Table 2.

The most frequent utterances for Naima labeled entities in the visual scene with VisObjP, and short answers accounted for most of Lily's utterances. Naima was a precocious talker whereas Lily, though good at answering Wh-questions, did not often initiate conversations verbally. This points to a possible shortcoming of our taxonomy: annotations were mainly driven by word or word-like utterances. Extra-linguistic cues were only used to guide category assignment to a somewhat intelligible verbal act and did not merit annotation on their own. This is a good first approximation. Nonetheless, the gestural actions initiating conversations also play an important role in language acquisition (Kelly, 2011). In the example below Lily uses pointing at pictures in a book and flipping pages as requests for labels:

(17) CHI: yyy . pho: ˈɪ sit: CHI turns the page back and points at it MOT: fish ! CHI: yyy . pho: ˈɛ MOT: turtle: . MOT: fish ! MOT: turtle . sit: CHI flipping page back and forth (Lily at 1;2.02)

Another reason we think this might be fruitful for the study of the early stages of child language is the high proportion of 'failed' (viz incomprehensible to the adult) utterances in the files we annotated; 39% of the utterances were unintelligible in Naima's files and this number goes even further up to 94% for Lily.

Repetition acknowledgments usually happened when a new label was provided by the caregivers. This is in line with the results of (Clark, 2007):

(18) sit: CHI crawling toward stuffed animal dog MOT: yeah there's puppy honey ! MOT: do you see puppy ? MOT: puppy's [: puppy is] over there . CHI: puppy . pho: ˈhʌbeɪ (Lily at 1;2.02)

For Naima, VisObjP Category becomes less frequent as she acquires new ways of referring to objects and moves to the two word stage, as suggested by the emergence of two word predication (e.g., "sleepy daddy") and other multi-word utterance denoting relations that are more sophisticated than simple labeling, at age 1;2.23 and 1;3.12, and also the increase in proportion of VisPred utterances. The 7.8% for VisPred at age 0;11.28 goes against this trend; but after taking a closer look at these utterances we discovered that this session was where Naima learned to say "down". The analysis of VisPred forms (Table 3) shows that the form diversity of VisPred utterances goes up with age. Similar analyses for other types in our taxonomy might prove useful for gaining more insight into children's developmental paths.

We calculated inter-annotator agreement scores using annotations done for a portion of file seven of Naima (approximately 25 minutes of conversation), by three other coders external to the project. The analysis of mismatches showed that RepAck

VisPred	0;11.28	1;0.28	1;1.25	1;2.23	1;3.12
# different forms	2	1	2	3	10
’down’	5	1	1	1	4
’good’	1		1		
’hot’				1	
’sleeping’				1	
’big’					7
’tall’					5
’heavy’					2
’stuck’					2
’off’					2
’dirty’					1
’on’					1
’kiss’					1
’clap’					1
total #	7	1	2	3	26

(ages)

Frequency of VisPred forms over time
Ages in years;months.days

Table 3: Frequency of VisPred for Naima

	r2	r3	r4
r1	86.66%, .83	89.19%, .85	80%, .60
r2		94.03%, .92	82.22%, .66
r3			82.05%, .78

Percent agreement %, Cohen’s kappa

Table 4: Inter-annotator agreement evaluation

and Imit categories are difficult to distinguish. Most mismatches however, were the result of the different choices by different coders for utterances to annotate, in the case of repetitive utterances. Removing these instances (i.e. including only utterances that both coders regarded as non repetitive), improved the scores significantly. These percent agreement, and Cohen’s kappa coefficient numbers are presented in Table 4.

5 Modelling parent/child interaction

5.1 Initial Assumptions

One of our main hypotheses is that parent (P)/early child (C) interaction involves mechanisms that, by and large, exist in regular adult/adult interaction. This is true for the principles underlying the coherence of such interaction and the types of utterances made by both parties. There is of course a basic asymmetry in that C is expected to adapt to P’s linguistic system and not vice versa (though there are parents who attempt the latter to a certain extent, as we will see.).

The main apparent differences derive from the fact that that much of the time C does not respond or responds in a fashion that is not comprehensible to P. Nor does P limit him/herself to uttering ‘comprehensible language’, in contrast to adult interaction where a basic presupposition exists of using ‘shared language’ (violating this is viewed, minimally, as arrogance.). However, given the flexibility of turn taking and the existence of dialogical repair mechanisms, this means that in the short term the conversation does not break down, whereas in the long term much positive evidence exploitable for learning gets produced.

5.2 Dialogue GameBoards

We use the dialogue framework KoS (see e.g. (Ginzburg and Fernández, 2010; Ginzburg, 2012) for details) as the framework for describing P/C interaction. On the approach developed in KoS, there is actually no single context—instead of a single context, analysis is formulated at a level of information states, one per conversational participant. This assumption is particularly useful for modelling an asymmetric type of interaction as here. The dialogue gameboard represents information that arises from publicized interactions. Its structure is given in (19)—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, though not simply questions qua semantic objects, but pairs of entities which we call *InfoStrucs*: a question and an antecedent sub-utterance (the *focus establishing constituent* (FEC) that partially specifies a subsequent focal utterance.[4]

(19) DGBType $=_{def}$

$$\begin{bmatrix} \text{spkr: Ind} \\ \text{addr: Ind} \\ \text{utt-time : Time} \\ \text{c-utt : addressing(spkr,addr,utt-time)} \\ \text{Facts : Set(Proposition)} \\ \text{Pending : list(locutionary Proposition)} \\ \text{Moves : list(locutionary Proposition)} \\ \text{QUD : poset(Infostruc)} \end{bmatrix}$$

DGBs are useful means of conceptualizing an adult’s public context in dialogue interaction. To

---

[4] On the whole, for current purposes one could restrict attention to QUDs consisting solely of questions. However, FECs potentially play a significant role in learning, as hinted in section 6.

what extent is this plausible for young children? It is plausible to assume that child representations of context are in several ways less detailed than adult ones, though some of the burden can be associated with incomplete mastery of the conversational rules we discuss shortly, rather than the representations. At the same time, given the evidence for common ground (Tomasello, 1988; Clark, 2013), for the ability to participate in simple games (amply demonstrated in Providence), and awareness of ‘topic continuity’ across utterances justifies the existence of (some notions corresponding to) FACTS, MOVES, and QUD respectively. We return to the issue of Pending below when we discuss metacommunicative interaction.

We make one modification to the DGB, one which is in any case also required for modelling adult/adult interaction. We introduce an extra field to FACTS which we dub *Vis(ual)Inf* of type *RecType* (cf. MSOA in (Grosz, 1977; Poesio and Rieser, 2011)). This represents the dialogue participant’s (view of) the visual situation and attended entities. The basic structure of this type is given in (20a). A concrete example is given in (20b): a visual situation involving a doll with spot on her head, where the spot is the attentional focus:

(20) a. $$\text{VisInf} = \left[\begin{array}{l} \text{VisSit : RecType} \\ \text{InAttention : Ind} \\ \text{c1 : member(InAttention,VisSit)} \end{array}\right]$$

b. $$\left[\begin{array}{l} \text{VisSit} : \left[\begin{array}{l} \text{x: Ind} \\ \text{c1 : doll(x)} \\ \text{y : Ind} \\ \text{c2 : head(y,x)} \\ \text{z : Ind} \\ \text{c3 : spot(z)} \wedge \text{On(z,y)} \end{array}\right] \\ \text{InAttention = VisSit.z : Ind} \\ \text{c1 : member(InAttention,VisSit)} \end{array}\right]$$

5.3 Conversational Rules

The basic units of change are mappings between dialogue gameboards that specify how one gameboard configuration can be modified into another on the basis of dialogue moves. We call a mapping between DGB types a *conversational rule*. The types specifying its domain and its range we dub, respectively, the *preconditions* and the *effects*, both of which are supertypes of DGBType.

An example of such a rule, taken from (Ginzburg, 2012), needed to analyze querying and assertion interaction is given in (21). QSPEC is what characterizes the contextual background of reactive queries and assertions. (21) says that if q is QUD–maximal, then subsequent to this either conversational participant—hence, the turn underspecification characterized by the type $TurnUnderspec$—may make a move constrained to be q–specific (i.e. either About or Influencing q).

(21) $$\begin{array}{l} \text{QSPEC} \\ \left[\begin{array}{l} \text{pre: } \left[\text{qud} = \left\langle \text{i, I} \right\rangle \text{: poset(InfoStruc)}\right] \\ \text{effects : TurnUnderspec} \\ \wedge_{merge} \left[\begin{array}{l} \text{r : AbSemObj} \\ \text{R: IllocRel} \\ \text{LatestMove =} \\ \text{R(spkr,addr,r) : IllocProp} \\ \text{c1 : Qspecific(r,i.q)} \end{array}\right] \end{array}\right] \end{array}$$

QSPEC highlights a feature of KoS’s dialogue semantics crucial for P/C interaction: the fact that a speaker can straightforwardly answer their own question. Such cases get handled because turn taking is abstracted away from querying: this allows either conversationalist to take the turn given the QUD-maximality of q.

Given how one sided interaction can be, also crucial are rules controlling downdating of questions without receiving responses (see section 8.3 in (Ginzburg, 2012)).

5.4 Metacommunicative Interaction

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 inferrable 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. This is also the basis for an account of interactive word learning (Macura, 2007; Larsson and Cooper, 2009).

In early child utterances, much of the time the adult does not react to incomprehensible utterances, but such reaction is certainly not rare.[5] At this stage the child does not initiate clarification interaction, but she clearly is sensitive to feedback about her utterances, both in terms of form

---

[5] *MOT: chew and swallow . *CHI: yyy . xpho: ba: *MOT: ba: . *MOT: what is ba ? *MOT: what d’you mean ba ? (from: naima2)

and meaning (Gallagher, 1977). This justifies the need for modifiable utterance representations akin to locutionary propositions—as we suggest below presumably incorporating probabilistic notions—and to some notion like PENDING. Given that the child does not initiate such interaction, how does it arise? We believe this is an instance of domain–specific reasoning about the dialogue, as we now explain.

5.5 Visually accessible chat

So far we have mentioned entirely domain general principles of interaction. In adult/adult interaction the activity type influences the interaction e.g. with respect to issues that arise without explicitly being introduced (cf. differences between conversations in a bakery, a train station, or among friends in a café.). Similar considerations apply here. As far as the parent is concerned, s/he confronts the following challenge—what to discuss with an interlocuter who, much of the time, does not respond in a comprehensible fashion and whose knowledge of language is very incomplete. The parent can talk about that which is visible and susceptible to linguistic description. Using the theory of conversational genres developed in (Larsson, 2002; Ginzburg, 2012), one may characterize this genre as *visually accessible chat* in which at any given point an inferrable issue is: *what word can one use to refer to the visually most prominent entity?*

This is the basis for our account of how the utterances (22(a),(b)) get the italicized readings.

(22) a. *MOT: should we comb her hair ? *MOT: with a comb ? *MOT(a): comb. ↦ *This entity can be referred to as a* ***comb*** (from: naima2)

b. sit: CHI reaching for MOT's microphone *MOT: that's a microphone . *CHI: microphone [?] . sit: CHI looking and pointing at MOT's microphone *MOT: mi:crophone . *MOT (b): a microphone . ↦ *This entity can be referred to as a* ***microphone*** (from: naima 4)

5.6 Initial child grammar

In this section we show how to formally characterize the utterance types which make up the taxonomy in section 3. For this purpose we use HPSG$_{TTR}$ (Ginzburg, 2012), a variant of the grammatical formalism Head–driven Phrase Structure Grammar (Sag et al., 2003). In specifying the child's grammar—a set (or type) of utterance types, we need to distinguish the comprehension grammar and production grammar. What we specify here are the production types (for interactions where the child is the speaker), but this is clearly distinct from the (presumably more extensive set/type of) comprehension types (for interactions where the child is the addressee).[6]

In terms of syntax, we follow the approach of (Ginzburg and Sag, 2000) to non-sentential utterances treating these as constructions with a single daughter, which constitutes the head.[7] Semantically, the context, represented within the field *dgb-params*, plays a crucial role via QUD, VisSit or Pending, providing the main predicate and/or the conversational move type.[8]

RepAck As with their use by adults, RepAck utterances are not straightforward to analyze: they can be viewed as bare acknowledgements ('an utterance containing this word was just uttered'.) or they can be viewed as singling out a word because the child is testing their pronunciation or understanding of the word. We propose that the conventional meaning of such utterances is essentially:[9] *child acknowledges that an utterance including the word word$_i$ happened.* (23) captures this by imposing segmental parallelism between a sub-utterance u1 of the prior (maximally pending) utterance and the AckRep utterance:[10]

---

[6] Indeed such a distinction probably needs to be drawn for the adult as well, e.g., to capture the difference between a carer of a given child and a random adult who interacts with child. But that is a somewhat more controversial case.

[7] In (Ginzburg and Sag, 2000) the category of these constructions was verbal, as by assumption this was the category of root utterances. We do not adopt this assumption here, which in any case is not ultimately tenable even for adult grammars, for a variety of interjections. We utilize a type *root*, whose explication we leave for another occasion. We are grateful to Joan Bresnan in conversation for alerting us to this issue.

[8] We present the types here in isolation. In a more detailed presentation one would extract some more general types and infer the 'leaves' of the type hierarchy using inheritance.

[9] (Clark, 2007) suggests that such utterances invariably involve recently acquired words. One could, in principle, include such a restriction in the construction description; it is unclear, however, whether carers are sensitive to this.

[10] Segmental parallelism is captured by imposing identity at the *type* level between u1 and the AckRep utterance. The ability to capture such parallelism distinguishes TTR from standard typed feature structure approaches to grammar.

(23)
$$\left[\begin{array}{l} \text{cat} = \text{root} : \text{syncat} \\ \text{dgb-params} : \\ \left[\begin{array}{l} \text{MaxPending} : \text{LocProp} \\ \text{u1} : \text{LocProp} \\ \text{c1}: \text{member(u1, MaxPending.sit.constits)} \\ \text{A} : \text{Ind} \\ \text{B} : \text{Ind} \\ \text{c1} : \text{address(A,B)} \end{array}\right] \\ \text{hd-dtr} : \left[\begin{array}{l} \text{tune} = \text{u1.sit-type.phon} : \text{Type} \\ \text{phon} : \textit{tune} \end{array}\right] \\ \wedge_{merge} \text{ sign} \\ \text{cont} = \text{Acknowledge(A,u1)} : \text{IllocProp} \end{array}\right]$$

VisObjP and VisPred In both cases the visually prominent entity plays a key role. For VisObjP it is simply that *entity InAttention has uttered word's descriptive property*, as in (24a). For VisPred the *property associated with uttered word is predicated of entity InAttention*, as in (24b).[11]

(24) a.
$$\left[\begin{array}{l} \text{cat} = \text{root} : \text{syncat} \\ \text{hd-dtr.cont} : \left[\begin{array}{l} \text{x} : \text{Ind} \\ \text{c1} : \text{P(x)} \end{array}\right] \\ \text{dgb-params} : \\ \left[\begin{array}{l} \text{v} : \left[\begin{array}{l} \text{VisSit} : \text{RecType} \\ \text{InAttention} = \text{hd-dtr.cont} : \text{Ind} \\ \text{member(InAttention,VisSit)} \end{array}\right] \end{array}\right] \\ \text{cont} = \text{P(hd-dtr.cont.x)} : \text{Prop} \end{array}\right]$$

b.
$$\left[\begin{array}{l} \text{cat} = \text{root} : \text{syncat} \\ \text{hd-dtr.cont} : (\left[\text{x} : \text{Ind}\right])\text{RecType} \\ \text{dgb-params} : \\ \left[\begin{array}{l} \text{v} : \left[\begin{array}{l} \text{VisSit} : \text{RecType} \\ \text{InAttention} : \text{Ind} \\ \text{member(InAttention,VisSit)} \end{array}\right] \end{array}\right] \\ \text{cont} = \text{hd-dtr.cont}(\left[\text{x=InAttention}\right]) : \text{Prop} \end{array}\right]$$

Short Answer In the adult grammar this is a much discussed construction (Morgan, 1973; Ginzburg and Sag, 2000; Merchant, 2004; Ginzburg, 2012), with controversy whether such constructions are underlyingly sentential or whether the resolution is effected semantically. In this case there is a clear preference for a semantically-based approach since it presupposes less syntactic knowledge for the child: semantically it involves predication—the question predicating of the fragment; it leaves open the issue whether the child is aware of parallelism between the interrogative whP and the fragment, as required in the adult case.[12]

(25)
$$\left[\begin{array}{l} \text{cat} = \text{root} : \text{syncat} \\ \text{hd-dtr.cont} : \left[\text{x} : \text{IND}\right] \\ \text{dgb-params} : \\ \left[\text{max-qud} : \text{UnaryWhQuestion}\right] \\ \text{cont} = \text{max-qud(hd-dtr.cont.x)} : \text{Prop} \end{array}\right]$$

ReqObj This is a class that is particularly interesting from a semantic point of view as this involves, arguably, the child expressing *non-referential* contents—a request for water does not involve asking for a specific portion, ditto when asking for more (wheat biscuit)—a well known puzzle in semantics first pointed out in (Quine, 1956) and satisfactorily solved in (Montague, 1974).[13] The TTR implementation of the latter analysis is based on (Cooper, 2005), in a non-higher order version proposed in (Ginzburg, 2012).[14,15] The type associated with ReqObj, given in (26a) uses the content of the word uttered by the child as the argument for the (illocutionary) Request predicate. This is exemplified for the utterance 'biscuit' in (26b)—the record type $\left[\begin{array}{l} \text{x} : \text{Ind} \\ \text{c1} : \text{biscuit(x)} \end{array}\right]$ represents the desire whose fulfillment the child requests—intuitively any witness for that type—an entity that is a biscuit—will do.

(26) a.
$$\left[\begin{array}{l} \text{cat} = \text{root} : \text{syncat} \\ \text{hd-dtr.cont} = \text{R} : \text{RecType} \\ \text{dgb-params} : \left[\begin{array}{l} \text{A} : \text{Ind} \\ \text{B} : \text{Ind} \\ \text{c1} : \text{address(A,B)} \end{array}\right] \\ \text{cont} = \text{Request(A,B, R)} : \text{IllocProp} \end{array}\right]$$

b. Child: biscuit $\mapsto$ Request(A,B,$\left[\begin{array}{l} \text{x} : \text{Ind} \\ \text{c1} : \text{biscuit(x)} \end{array}\right]$)

---

[11] To avoid notational clutter, we omit the assertoric illocutionary force associated with these utterances and with short answers.

[12] In (Ginzburg and Sag, 2000; Ginzburg, 2012) this parallelism is captured by constraining the category of the head daughter to be identical to the focus establishing constituent, in this case the sub-utterance corresponding to the *wh*-phrase.

[13] For reasons of space we do not discuss the REQ event type here. This would involve a futurate propositional entity such as an *outcome* (Ginzburg and Sag, 2000), presumably including the child as agent.

[14] The witnessing conditions of the record type that fills the object argument role seem to describe well the fulfillment conditions of a desire.

[15] Of course, as Dimitra Kolliakou (p.c.) has pointed out to us, one could argue that the child does not utilize non-referential contents at this stage, exploiting an image of a recent token or some such.

6 The emergence of the initial grammar

Our ultimate aim is not merely to describe the mechanisms of a single period of child utterances in dialogue, but to develop a theory that can accurately describe the transitions between distinct phases. Such a theory should, to the extent possible, explicate this on the basis of interaction between parent and child, given plausible priors and general learning principles. Developing such a theory has been a long term aim since at least (Bruner, 1981).

Here we sketch quite informally for two of the utterance types discussed in section 5.6 some components of an interaction–oriented theory accounting for their emergence.

RepAck once the child understands that she is expected to participate using words, that turns are assigned to her, and can chunk an utterance, those responses of hers that resemble recently uttered words get differentially positive feedback.[16] RepAcks serve as a probe for the child's ability to imitate correctly and with appropriately fast timing. Feedback from the adult causes the child to adjust her hypotheses about a new word. For such hypothesis adjustment using probabilistic representations in TTR see (Cooper et al., 2013).

Short answer The fundamental problem here is to learn the answerhood relation holding between interrogative utterances and (certain classses of) subsequent utterances. The child gets significant data on this from the parent who responds to the parent's own queries—the child receives evidence for several possible answers to a question and several forms, sentential and non-sentential. But this, in turn, presupposes that the child has some means of classifying utterances as wh–interrogative. Morphosyntactically, we can assume this as a prior. But the issue that remains is distinguishing the meaning of different wh–words, at this stage *where, what, who*, as well as combining these with predicates. For the former, we hypothesize this can be done on the basis of utterance bigrams linking *where*–utterances with deictic gestures/locative utterances to entities in the visual field, in contrast to *what*–utterances which are differentially linked to utterances supplying attributes.

7 Conclusions and Further Work

In this paper we develop a taxonomy of early child utterances that allows the fine structure of the semantic content of such utterances to be represented, thereby remedying problems for existing classifications. We offer a formal analysis of such utterances in the frameworks of KoS and HPSG$_{TTR}$. This requires spelling out the dialogue context and interaction since such utterances are strongly context dependent. We also provide a brief sketch of how two classes of such utterances could be acquired interactively.

In ongoing work, we are refining the taxonomy to incorporate gesture and to scale up to later, more complex utterances. We also intend to implement a learning algorithm which will allow us to experimentally test the interactive account of acquisition of certain early utterance types, above all short answers.

References

John Langshaw Austin. 1975. *How to Do Things with Words*. Oxford University Press.

Elizabeth Bates. 1976. *Language and context: The acquisition of pragmatics*, volume 13. Academic Press New York.

Lois Bloom. 1967. Language Development: Form and Function in Emerging Grammars. November.

Jerome Bruner. 1975. The ontogenesis of speech acts. *Journal of Child Language*, 2(01):1–19, September.

Jerome Bruner. 1981. The social context of language acquisition. *Language & Communication*, 1(2):155–178.

Eve V. Clark. 2007. Young children's uptake of new words in conversation. *Language in Society*, 36:157–182.

Eve Clark. 2013. Common ground. In Brian MacWhinney and William O'Grady, editors, *The Handbook of Language Emergence*. Wiley-Blackwell, London.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2013. Probabilistic type theory and natural language semantics. *Under Review*, xx:xx.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(4):333–362.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2):137–173.

---

[16] *MOT: that baby has a bottle , did you notice that ? *CHI: bottle . *MOT: yeah , baby has a bottle. (from: Naima03)

John Dore. 1974. A Pragmatic Description of Early Language Development. *Journal of psycholinguistic Research*, 3(4):343–350.

John Dore. 1975. Holophrases, speech acts and language universals. *Journal of Child Language*, 2(01), September.

Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances: A corpus-based study. *Traitement automatique des languages*, 43(2):13–42.

Tanya M Gallagher. 1977. Revision behaviors in the speech of normal children developing language. *Journal of Speech, Language and Hearing Research*, 20(2):303.

Jonathan Ginzburg and Raquel Fernández. 2010. Computational models of dialogue. In Alex Clark, Chris Fox, and Shalom Lappin, editors, *Handbook of Computational Linguistics and Natural Language*, Oxford. Blackwell.

Jonathan Ginzburg and Ivan A. Sag. 2000. *Interrogative Investigations: the form, meaning and use of English Interrogatives*. Number 123 in CSLI Lecture Notes. CSLI Publications, Stanford: California.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Barbara Grosz. 1977. *The Representation and Use of Focus in Dialogue Understanding*. Ph.D. thesis, Stanford University.

Ray Jackendoff and Eva Wittenberg. 2014. What you can say without syntax: A hierarchy of grammatical complexity. In Fritz Newmeyer and Laurel Preston, editors, *Grammatical Complexity*. Oxford University Press, New York, NY.

Barbara Kelly. 2011. A new look at redundancy in children's gesture and word combinations. In Inbal Arnon and Eve Clark, editors, *Experience, Variation and Generalization: Learning a First Language*. John Benjamins, Amsterdam.

Staffan Larsson and Robin Cooper. 2009. Towards a formal view of corrective feedback. In *Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition*, Athens.

Staffan Larsson. 2002. *Issue based Dialogue Management*. Ph.D. thesis, Gothenburg University.

Zoran Macura. 2007. *Metacommunication and Lexical Acquisition in a Primitive Foraging Environment*. Ph.D. thesis, King's College, London.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, 26(4):657–657.

Jason Merchant. 2004. Fragments. *Linguistics and Philosophy*, 27:661–738.

Richard Montague. 1974. The proper treatment of quantification in ordinary english. In Richmond Thomason, editor, *Formal Philosophy*. Yale UP, New Haven.

Jerry Morgan. 1973. Sentence fragments and the notion 'sentence'. In B Kachru, editor, *Issues in Linguistics: papers in honour of Henry and Rene Kahane*. UIP.

Anat Ninio and Jerome Bruner. 1978. The achievement and antecedents of labelling. *Journal of Child Language*, 5(01):1–15, September.

Anat Ninio and Catherine Snow. 1988. Language Acquisition Through Language Use: The Functional Sources of Children's Early Utterances. *Y. Levy, I. Schlesinger, & M.D.S. Braine (Eds.), Categories and processes in language acquisition*, pages 11–30.

Anat Ninio and Polly Wheeler. 1986. A manual for classifying verbal communicative acts in mother-infant interaction. *ERIC*, 3:1–12.

Anat Ninio, Catherine E. Snow, Barbara A. Pan, and Pamela R. Rollins. 1994. Classifying communicative acts in children's interactions. *Journal of Communication Disorders*, 27.

Anat Ninio. 1992. The Relation of Children's Single Word Utterances to Single Word Utterances in the Input*. *Journal of Child Language*.

Massimo Poesio and Hannes Rieser. 2011. An incremental model of anaphora and reference resolution based on resource situations. *Dialogue and Discourse*, 1:235–277.

Willard V Quine. 1956. Quantifiers and propositional attitudes. *the Journal of Philosophy*, 53(5):177–187.

Ivan A. Sag, Thomas Wasow, and Emily Bender. 2003. *Syntactic Theory: A Formal Introduction (second edition)*. CSLI, Stanford.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*, volume 0. Cambridge University Press.

Marilyn Shatz. 1979. How to Do Things by Asking : Form-Function Pairings in Mothers' Questions and Their Relation to Children's Responses. *Child Development*, 50(4):1093–1099.

Catherine E. Snow. 1972. Mothers' Speech to Children Learning Language. *Child development*, 43(2):549–565.

Michael Tomasello. 1988. The role of joint attentional processes in early language development. *Language Sciences*, 10(1):69–88, January.

English reverse prosody in yes-no responses

Daniel Goodhue
Linguistics Department
McGill University
Montréal, Québec, CANADA
daniel.goodhue@mail.mcgill.ca

James Pickett
Department of Linguistics and English Language
The University of Manchester
Manchester, UK
jamesmtpickett@gmail.com

Michael Wagner
Linguistics Department
McGill University
Montréal, Québec, CANADA
chael@mcgill.ca

Abstract

In English, polar particles *yes* and *no* are ambiguous when used to respond to negative declaratives and interrogatives. This paper reports on a production experiment that elicited the intonation contours speakers use when responding to negative declaratives. We found that speakers most frequently use the Contradiction Contour when reversing, and they use declarative intonation when confirming, regardless of the particular polar particle used. Therefore prosody could disambiguate what is an otherwise ambiguous move in a dialogue.

1 Introduction

English polar particles *yes/yeah* and *no* are ambiguous when responding to negative declaratives/interrogatives, whereas these polar particles are unambiguous when responding to positive declaratives/interrogatives. (Cf. Cooper and Ginzburg, 2011a; Farkas and Roelofsen, 2013; Holmberg, 2012; Kramer and Rawlins, 2009; Krifka, 2013)

(1) A: Matt called
A: Did Matt call?
A: Matt called?
a. B: Yes/Yeah, Matt called
b. B: No, Matt did not call
c. B: # Yes/Yeah, Matt did not call
d. B: # No, Matt called

(1) shows possible responses to positive declaratives and interrogatives as reported in the literature. *Yes/yeah* and *no* can be uttered with or without the following sentences in (1-a) and (1-b) respectively. That the polar particles are unambiguous is reflected by the infelicity of (1-c) and (1-d), as opposed to the following paradigm in (2) where all responses are felicitous.[1]

(2) A: Matt didn't call
A: Did Matt not call?
A: Matt didn't call?
a. B: Yes/Yeah, Matt called
b. B: No, Matt called
c. B: Yes/Yeah, Matt did not call
d. B: No, Matt did not call

(2) shows possible responses to negative declaratives and interrogatives. The responses from (2-a) through (2-d) are all acceptable (but see below). Therefore, if a person says only *yeah* or *no* in response to (2), it is ambiguous whether that person means that Matt called or that he did not.

This paper reports a production experiment that we believe makes contributions to three questions about (2) that have remained somewhat controversial in the literature: **1)** Do the particles and the sentences in the responses in (2) bear particular intonational contours, if so which, and on which responses? **2)** Are some responses in (2) more natural than others? **3)** Are polar particle responses infelicitous if they are not accompanied by sentences or sentence fragments, like those in (2)?

---

[1] However, responses like (1-d) may be acceptable in particular contexts where speaker A questions some presupposition that is so obviously true that it makes the negative answer salient/produces a negative bias. Then speaker B may be licensed to say "No" followed by a positive sentence. Experimental testing may be required to establish this. E.g.:

(i) A: Guess what? I won tickets to see Justin Bieber.
B: [does not react]
A: Do you know who Justin Bieber is?
B: No, I know who Justin Bieber is. I just don't care.

Regarding **1)**, Cooper and Ginzburg (2011a) report that to the extent that *no* is ambiguous in contexts like (2), the reverse meaning in (2-b) will bear a distinct rise fall tune, whereas the confirm meaning in (2-d) is most naturally associated with a fall. Farkas and Roelofsen (2013) claim that responses similar to (2-a) and (2-b) except that they contain sentence fragments (e.g. *Yes/No, he did*) must bear stress on the auxiliary verb, *did*. Alternatively, they claim that speakers can use what they call "Smart Aleck" intonation which rises on the particle and falls on the auxiliary. Krifka (2013) claims that responses like (2-a) and (2-b) require a rejecting accent when responding to a negative assertion, though he doesn't describe what the accent is. These accounts do not offer experimental evidence for the intonations they discuss. The experiment reported here contributes new information regarding question **1)** by showing that the responses in (2-a) and (2-b), which reverse the preceding utterance by having opposite polarity from it, most frequently bear the Contradiction Contour (Liberman and Sag, 1974) on either the polar particle, the following sentence or both. Prior literature has not discussed the use of the Contradiction Contour in contexts like (2). We also found that confirming responses such as (2-c) and (2-d) almost always bear declarative falling intonation on either the polar particle, the following sentence or both.

Regarding **2)**, Krifka (2013) uses an optimality theoretic framework to argue that the preference of responses to (2) are ranked in the following order from most to least acceptable: (2-d), (2-c), (2-a), (2-b).[2] Our experiment provides a different answer to **2)** in the form of naturalness judgments given by participants that reveal that all responses in (2) are judged natural with the exception of (2-c) which is somewhat degraded.[3] [4]

Regarding **3)**, both Farkas and Roelofsen (2013) and Krifka (2013) claim that reverse responses like (2-a) and (2-b) must occur with following sentences. We do not answer question **3)** here, though the asymmetry of intonational contours we found on the polar particles suggests that speakers may be able to distinguish the meanings of the responses in (2) in the absence of following sentences on the basis of intonation. Future experimentation is required to establish this.

In section 2, we will briefly discuss Krifka's (2013, to appear) theories of polar particles and reversing moves in conversations. Then we will characterize the Contradiction Contour phonologically and semantically, and discuss its relation to Krifka's account. In section 3, the methods of the experiment are described. In section 4, the experimental results are discussed. In 5, we conclude and discuss future directions.

2 REJECT and the Contradiction Contour

In this section we will briefly describe Krifka's (2013) theory of polar particles and how it is linked to the REJECT operator, which Krifka (to appear) claims is sometimes encoded by English "protest prosody". Then we will describe the Contradiction Contour phonologically and semantically, and discuss it's connection to REJECT.

Krifka (2013) analyzes polar particles as anaphoric expressions that refer to some antecedent in the discourse. He compares them to other propositional anaphora like *that*.

(3) A: Two plus two isn't five
$[_{\text{NegP-}\phi}$ NEG $[_{\text{TP-}\psi}$ 2+2 is 5$]]$
a. B: Everyone knows *that* (i.e. ϕ)
b. B: *That* (i.e. ψ) would be a contradiction

Propositional anaphora find two possible antecedent discourse referents in negative phrases, like (3). One propositional discourse referent, ϕ, is made available by the NegP, and another, ψ, is produced by the TP. *That* can refer to either proposition, as seen in (3-a) and (3-b).

Krifka proposes that *yes* picks up a salient propositional discourse referent and asserts it. *No* picks up a salient discourse referent and negates it.

(4) A: $[_{\text{TP-}\psi}$ Maxine arrived on time$]$
A: $[_{\text{CP}}$ Did $[_{\text{TP-}\psi}$ Maxine arrive on time$]]$
a. B: Yes = ASSERT(ψ)
b. B: No = ASSERT($\neg\psi$)

---

[2] Krifka notes that this ranking is context dependent.

[3] Brasoveanu et al. (2011) found that speakers prefer (2-d) over (2-c) when the subject of the sentence is a referential NP (e.g. *Matt/he*). When the subject is shifted to an upward monotone quantifier (e.g. *some X*), the preference disappears, and the preference flips if the subject is a downward monotone or non-monotone quantifier (e.g. *at most X* or *exactly X* respectively). Only referential NPs were used in our experiment.

[4] Cf. Cooper and Ginzburg (2011b) who report a corpus study that shows that positive polar interrogatives are more likely to elicit a positive response whereas negative polar interrogatives are more likely to elicit a negative response.

In (4), we see that positive assertions and interrogatives only make one propositional discourse referent available as an antecedent, ψ. Therefore, *yes* can only assert ψ and *no* can only negate it, which captures the data in (1) above.

In (3), A's negative statement produced two discourse referents. The same happens in (5):

(5) A: [$_{\text{NegP-}\phi}$ NEG [$_{\text{TP-}\psi}$ Maxine arrived on time]]
A: [$_{\text{CP}}$ Did [$_{\text{NegP-}\phi}$ NEG [$_{\text{TP-}\psi}$ Maxine arrive on time]]]
a. B: Yes = ASSERT(ψ)
b. B: Yes = ASSERT(ϕ)
c. B: No = ASSERT($\neg\psi \approx \phi$)
d. B: No = ASSERT($\neg\phi \approx \psi$)

In (5), A utters a negative declarative *Maxine didn't arrive on time*, or a negative interrogative *Did Maxine not arrive on time?*. Each utterance makes two propositional discourse referents available: ϕ is produced by NegP, ψ is produced by TP. In (5-a), *yes* picks up ψ and asserts it, while in (5-b), *yes* picks up ϕ and asserts it. In (5-c), *no* picks up ψ and negates it (which approximates ϕ), while in (5-d), *no* picks up ϕ and negates it (which approximates $\neg\neg\psi$, which in turn approximates ψ).[5] Therefore, Krifka's account captures the ambiguity seen in (2).

Krifka's (to appear) theory of conversation states that speakers attempt to add a proposition to the common ground when they utter any kind of declarative, including rising declaratives (e.g. *Dave called*, *Dave called?*, *Dave didn't call*, and *Dave didn't call?*) and negative interrogatives (e.g. *Did Dave not call?*). If an interlocutor wants to deny the addition of one of these propositions to the common ground (e.g. by uttering *no* or *I don't believe that*), a REJECT operator is required to remove the first proposition from the common ground. Otherwise both the initial proposition and the denial of that proposition would be in the common ground, creating an inconsistent context set. Although REJECT is obviously not encoded by a single expression, it is encoded lexically in polar particles of some languages (e.g. *si* in French and *doch* in German). When denying or reversing negative declaratives and interrogatives in English, Krifka claims REJECT is encoded as "protest prosody". Therefore, Krifka's theory predicts protest prosody to appear in responses like (5-a) and (5-d) where B's response contradicts A's initial utterance.

As mentioned above, Krifka (2013, to appear) does not further characterize the protest prosody/rejecting accent of English. The goal of this paper is to characterize the prosodic intonation English speakers use when reversing and when confirming preceding questions, more particularly uninverted negative questions with a final rise ('rising declaratives'). We found that English speakers frequently use the Contradiction Contour (CC) on the polar particle and/or the following sentence when uttering a positive proposition that reverses the negative proposition of the preceding utterance. Moreover, speakers rarely use the CC when confirming the negative proposition of the prior utterance, and they judged such utterances unnatural.

The CC has been described by Liberman and Sag (1974) as an utterance wide contour that has an initial rise, with a fall across most of the utterance followed by an utterance final rise. Two separate instances of the CC can be found in figure 1 below: the first on "No", the second on "I'm a friend of Jenny's". The second utterance most clearly illustrates Liberman and Sag's description. They do not discuss what form the CC would take when it appears on a monosyllabic utterance such as "No". We found that in such cases the CC falls initially before meeting the utterance final rise (see the first pitch track in figure 1). This is perhaps unsurprising since, according to Ladd (1980), the CC places a low pitch accent on the nuclear stress of the utterance, with a high falling tone preceding the nucleus, which is what we see on "No". Liberman and Sag (1974) characterize the meaning of the CC as follows: "We find that this contour is appropriate (although of course optional) just when the speaker is using the utterance that bears it to contradict–he may contradict what has just been said by another, he may contradict some assumption or implication of what has been said or done by another, or he may contradict himself." (pg 421)

From the observations of the CC's distribution in our experiment and the observations of Liberman and Sag, we argue that the CC is a prosodic contour that is felicitous on an utterance of the

---

[5] Cf. Cooper and Ginzburg (2011a) for a different approach that analyzes $\neg\neg\psi$ and ψ as truth-conditionally equivalent, but not identical, propositions. Krifka assumes a classical logic framework in which $\neg\neg\psi$ and ψ are equivalent.

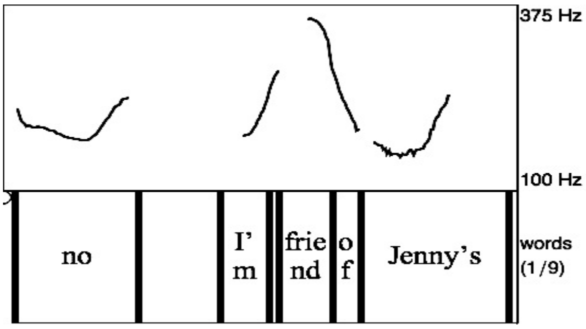


Figure 1: *F0 pitch track of the contradiction contour appearing twice: once on "No", and once on "I'm a friend of Jenny's".*

proposition ϕ if and only if the proposition $\neg\phi$ is salient in the context. This characterization of the CC is formalized in (6).

(6) $[\![\mathrm{CC}]\!]^{c} = \lambda \mathrm{P}_{\langle s,t\rangle}$: $\neg$P is salient in c. P

Notice that the negation could be swapped from one proposition to the other here. I.e. the CC can be uttered on a proposition $\neg\phi$ if ϕ is salient in the context. All that is needed to license the CC is a salient proposition that is incompatible with the CC proposition. An explanation of how one proposition is recognized to be the negation of another is beyond the scope of this paper.[6]

Our analysis of the CC is similar to Krifka's proposal in that there is a REJECT operator whose presence can be encoded by prosody. Our approach diverges from Krifka's, however, in at least one way: We attribute the meaning of the REJECT operator directly to the contour, and assume that in the absence of the contour there is no REJECT operator. Therefore the distribution of the CC-REJECT is similar, but not identical, to the distribution of Krifka's REJECT. For example, Krifka's REJECT operator is necessary for any denying move in a dialogue, whereas the CC appears to be optional, as Liberman and Sag (1974) already point out. Moreover the distribution of Krifka's REJECT is not as restricted as that of the CC in (6), as evidenced by the fact that his REJECT appears when disbelief in a proposition is expressed (without its negation being salient)—a move that does not license the CC. In section 4 below, we will show how our characterization of the CC accounts for the contour's distribution.

The goal of of our experiment is to capture which intonations English speakers use when responding to negative rising declaratives, and to obtain naturalness judgments about the responses participants were asked to produce. We expected that there would be an asymmetry between the intonations used in the Reverse conditions and those used in the Confirm conditions, a prediction shared to varying degrees by researchers who have studied English response particles, including Cooper and Ginzburg, Farkas and Roelofsen, Holmberg, Kramer and Rawlins, and Krifka.

3 Methods

The participants were 22 North American English speakers, mostly undergraduate students. There were six items, each comprised of six conditions, four test-conditions with negative rising declaratives and two additional conditions which we will not report on in this paper for reasons of space. The trials were pseudo-randomized so that participants never saw the same condition twice in a row, and trials from the same item were organized into different blocks to maximize their distance.

Participants were presented with a context story on a computer screen. After they had read it, they pressed a key to hear a question through headphones. Then they pressed a key to start recording their response to the question. Participants were given a script to use for responding. Then participants were asked to judge the naturalness of the response on a scale from 1 to 5. Below are example contexts, questions and responses for a reverse response and a confirm response.

(7) Reverse context:
You are at home eating lunch. After several days of rain it's warm and sunny, and you are planning to go to the park after you finish eating. Your new roommate walks in and asks if you want to go to the movies with him this afternoon. You like movies and want to see a film thats currently at the theater, but not today because the weather is so nice you want to take advantage of it by being outside. When you tell him you'll pass, your new roommate asks:
Q: You don't like movies?
A: No ___ I like movies.

(8) Confirm context:

---

[6] Cf. Farkas and Roelofsen (2012) for an account in terms of complementary sets of possible worlds, and Cooper and Ginzburg (2011a,b) for an account in terms of Type Theory with Records.

> You are at home eating lunch. After several days of warmth and sun a storm has moved in and it's raining. As you eat, you are trying to figure out what you will do with your afternoon. Your new roommate walks in and asks if you want to go to the movies with him. This would be a good solution except that you hate movies and prefer to spend your time reading or talking with friends. When you tell him you'll pass, your new roommate asks:
> Q: You don't like movies?
> A: No ___ I don't like movies.

Participants were instructed to treat the "___" in the responses as a pause between the polar particle and the following sentence. This was done to ensure that participants produced an intonation contour unique to the polar particle rather than producing a single contour across the entire utterance. Below the four conditions with negative questions in the context:.

(9) Experimental conditions
- a. **Question**: You don't like movies?
- b. **Yes-Reverse**: Yeah, I like movies.
- c. **No-Reverse**: No, I like movies.
- d. **Yes-Confirm**: Yeah, I don't like movies.
- e. **No-Confirm**: No, I don't like movies.

Each token was categorized for the intonation that appeared on the polar particle and again for the intonation that appeared on the following sentence. Intonations produced by participants were separated into four categories: the contradiction contour (CC) was described in section 2; declarative falling intonation (Declarative) has been identified by Pierrehumbert and Hirschberg (1990) as H* L L% in ToBI transcription; rise fall intonation (RiseFall), which is probably a variation of declarative intonation, and which rises to a high peak on the nuclear stress of the sentence and then falls. We excluded a small number of utterances which seemed to carry a different contour (such as the so-called rise-fall-rise contour) or which we couldn't easily classify along this scheme. Contra Liberman and Sag (1974), we consider the CC and the rise-fall-rise to be two (of at least four) distinct contours in English. We agree with Ladd (1980) that they can be distinguished in terms of the location of the rise (preceding the nucleus in the CC, on the nucleus in rise-fall-rise).

4 Results and discussion

In this section, we present the results of a production study that recorded English speakers' responses to negative rising declaratives. We show how the results can be explained by our characterization of the CC above in section 2. Finally we discuss participants' naturalness judgments of the responses they produced.

Figure 2 represents the amount that each intonational category was produced on the polar particles *yeah* and *no* in response to negative rising declaratives like *You don't like movies?*. The y axis shows the percent that each category was produced per condition. The x axis indicates which condition each bar refers to. The percentages for each category are stacked into a single bar for each condition. From bottom to top: the CC is dark grey, Declarative fall is medium grey, and RiseFall is in light grey.

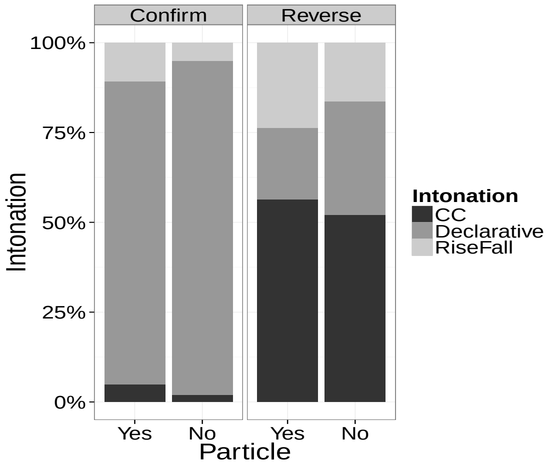


Figure 2: *Frequency of particle intonation per condition as a percentage*

Figure 2 and table 1 (below) show that the CC appears on 56% of *yeah* particles and 52% of *no* particles in the Reverse conditions, and that it appears on 5% and 2% of *yeah* and *no* particles respectively in the Confirm conditions. Given our characterization of the CC above in section 2, this is unsurprising. Under our analysis, the CC can appear on a proposition only if the negation of that proposition is salient in the context. The proper context for the CC is created in the Reverse conditions because the questioner made $\neg\phi$ salient (e.g. *You don't like movies?* = $\neg\phi$), so the participant

Table 1: Percent of intonation response on polar particle by condition.

Meaning	Particle	CC	Dec	RF
Confirm	Yes	5%	84%	11%
	No	2%	93%	5%
Reverse	Yes	56%	20%	24%
	No	52%	32%	16%

is licensed to utter ϕ with the CC (e.g. *Yeah/No, I like movies* = ϕ). Moreover, the proper context for the CC is not created in the Confirm conditions because the questioner and the participant utter the same proposition $\neg\phi$ (e.g. *The participant doesn't like movies* = $\neg\phi$).

Figure 2 and table 1 further reveal that the polar particles in the Confirm conditions bore declarative intonation 84% for *yeah* and 93% for *no*. In Reverse conditions, polar particles bore declarative intonation 20% for *yeah* and 32% for *no*, which shows that, although the CC is the most frequent contour when reversing, declarative intonation is still a licit contour when reversing. Finally, RiseFall intonation was produced 24% for *yeah* and 16% for *no*. Recall that RiseFall is a special instance of declarative intonation that contains a high peak.

We analyzed the data by coding a binary factor for whether or not the CC was used, and conducted a mixed model logistic regression with Particle ('yes' or 'no) and Reversal ('reverse', 'confirm') and their interaction as fixed factors, and participant and item as random effects that included slopes for the fixed factors and their interaction. We found a significant main effect of Reversal ($z = 6.4, p < 0.001$), and no main effect of Particle ($z = -0.57, p < 0.32$) and no interaction between Reversal and Particle ($z = -0.82, p < 0.47$). In other words, the choice between 'yes' and 'no' had no influence on the choice between the CC intonation and other options. We used a mixed model logistic regression analysis over alternatives such as ANOVA because ANOVA is inadequate in the analysis of proportions (Jaeger 2008), and mixed model logistic regression allows the researcher to control for item and participant random effects at the same time. For discussion on why other methods are problematic see Barr et al. (2013) and Baayen (2008).

Figure 3 represents the amount that each intonational category was produced on the sentences following the polar particles in response to negative rising declaratives like *You don't like movies?*.

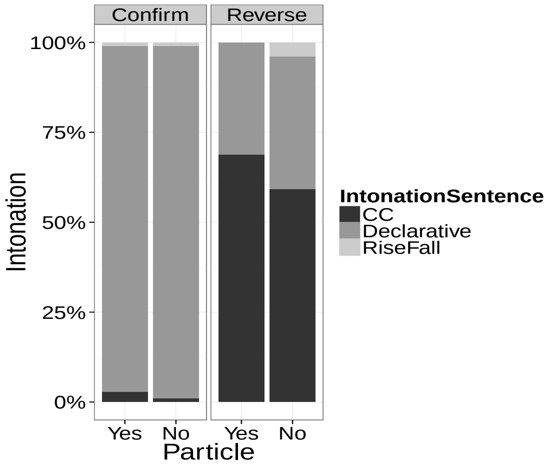


Figure 3: *Frequency of sentence intonation per condition as a percentage*

Table 2: Percent of intonation response on sentence by condition.

Meaning	Particle	CC	Dec	RF
Confirm	Yes	3%	96%	1%
	No	1%	98%	1%
Reverse	Yes	69%	31%	0%
	No	59%	37%	4%

Figure 3 and table 2 show that the CC appears in the Reverse conditions on 69% of the sentences following *yeah*, and on 59% of the sentences following *no*. The CC appeared in the Confirm conditions on 3% of sentences following *yeah*, and 1% of sentences following *no*. This is unsurprising for the same reasons that it was unsurprising for figure 2. However in figure 3 the CC accounts for an even greater proportion of the intonations in the Reverse conditions than in figure 2.

Figure 3 and table 2 reports also show that, in the Confirm conditions, declarative intonation was produced on 96% of sentences following *yeah* and 98% of sentences following *no*. In Reverse conditions, declarative intonation was produced on 31% of sentences following *yeah* and on 37% of sentences following *no*.

Again, we coded choice of CC in a binary factor and fitted the same type of mixed model for the choice of sentence contour. We found a significant main effect of Reversal ($z = 2.3, p < 0.02$), and no main effect of Particle ($z = 0.23, p < 0.81$) and no interaction between Reversal and Particle ($z = -0.30, p < 0.77$). In other words, the choice between 'yes' and 'no' had no influence on the choice between the CC intonation and other options on the sentence.

So both on the particle and on the sentence, speakers were likely to use the CC in the reverse condition but not in the confirm condition. It is interesting to show the break down of how well the intonations on the two constituents correlated. Figure 4 shows which intonation participants produced on the following sentence dependent on whether they produced the CC on the preceding polar particle. The x axis represents whether or not participants used the CC on the polar particle. The y axis represents what percentage the participants produced the CC vs. non-CC on the following sentence.

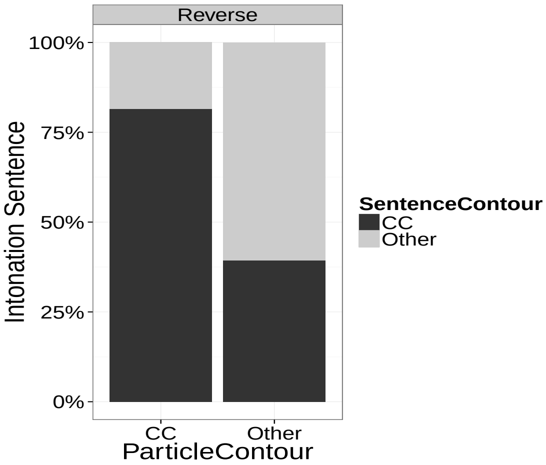


Figure 4: *Sentence intonation within particle intonation*

Of primary interest in figure 4 is what participants did when they produced a non-CC intonation on the polar particle in the reverse condition (the bar on the right): participants produced the CC on the sentence following a non-CC polar particle on 39% of utterances. This means that in over a third of reverse utterances that did not bear the CC on the polar particle in figure 2 above, the participant went on to produce the CC on the following sentence. Since the CC was used very rarely on the polar particle in the confirm condition, those data points are not represented in figure 4.

It seems reasonable to assume that a reversal is encoded in a response when the CC is placed on *either* constituent. Figure 5 shows the percentage of responses bearing the CC in the confirm and reverse condition. If the participant produced the CC either on the particle or the following sentence or both, we count the utterance as bearing the CC (dark grey); if it didn't bear the CC on the particle nor on the sentence, then that observation is counted No CC (light grey).

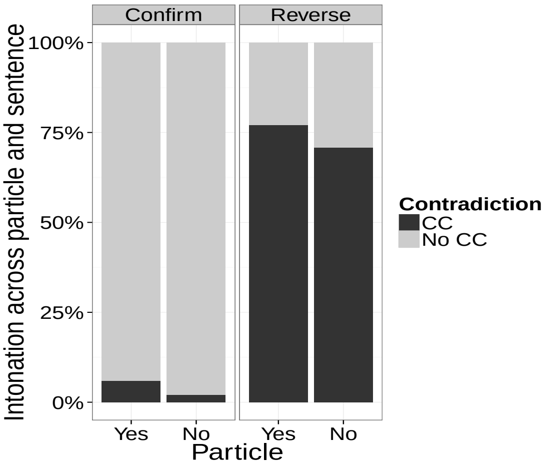


Figure 5: *Intonation across particle and sentence*

The CC appears at least once in 76.9% of responses in yes-reverse (third bar from left) and in 71.4% of responses in no-reverse (fourth bar from left). This means that participants produced the CC contour either on particle or sentence 74% of the time in the reverse condition. Moreover, participants only produced the CC 4% of the time in the confirm condition. Therefore not only is the presence of the CC strongly correlated with reversal of the salient negative proposition, but the absence of the CC is strongly correlated with confirmation of the negative proposition made salient by the question.

In section 1 above we posed the following question: **1)** Do the particles and the sentences in responses to negative utterances bear particular intonational contours, if so which, and on which responses? The answer suggested by the data in this section is that reverse responses (responses with opposite polarity from the negative declarative they respond to) bear the CC on the polar particle or the following sentence 74% of the time, but confirm responses (responses with the same polarity as the negative declarative they respond to) do so very rarely. We assume that the small number of occurrences that we found might be inflated because overall the CC-conducive contexts were very frequent in the experiment and there may have been some persistence of intonational uses across trials.

4.1 Naturalness judgments

Figure 6 shows participants' judgments of the naturalness of the responses they were asked for. All conditions show a median naturalness rating of 4, except for *yeah* confirming responses, which received a median naturalness judgment of 3.

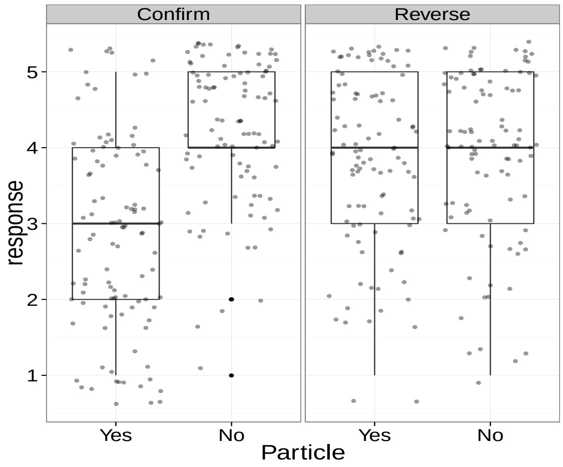


Figure 6: *Naturalness judgment on a scale of 1 to 5*

The results are interesting because they suggests that speakers don't have strong preferences for using *yeah* over *no* or vice versa when reversing negative declaratives, contrary to what is expected according to Krifka's (2013) optimality theoretic account of speaker preferences for certain responses over others discussed above in section 1. Moreover, *yeah* for confirming is judged somewhat degraded compared to the other responses while Krifka ranks this response as second most acceptable. In Farkas and Roelofsen 2013 the authors report that *yeah* in some dialects of English can only confirm a prior utterance (pg 23). Since our experiment used *yeah* to the exclusion of *yes*, one would then expect *yeah* as a reverse response to be judged unnatural and as a confirm response to be judged natural, contrary to our data. Perhaps the North American English speakers tested in Montreal, QC (a mix of Canadians and Americans) do not speak the relevant dialect. These naturalness data then start to reveal an answer to question **2)** posed in section 1: Are some responses to negative utterances more natural than others? The answer suggested by our data is, they are all fairly natural, although *Yeah, I don't like movies* slightly less so.

5 Conclusion

This paper reported on a production experiment investigating the prosodic tunes English speakers produce when responding to negative questions. The experiment showed that in lab contexts, when the response reverses the negative bias of the question, speakers produce the Contradiction Contour (CC) 74% of the time, so it is by far the preferred sentence contour in this context. When the response confirms the negative bias of the question, speakers produce the CC a negligible amount, but instead overwhelmingly produce declarative intonation. Gaining an understanding about when particular contour is preferred/dispreferred is an important step in figuring out what its semantic and pragmatic content is (For a similar attempt at characterizing the contexts in which speakers produce/avoid the rise-fall-rise contour see Wagner et al., 2013). The particular polar particle produced in the response (e.g. *yeah* vs. *no*) had no effect on the intonation observed.

In section 1 above, we identified three questions of interest regarding the ambiguity of polar particles *yes* and *no* when responding to negative declaratives and interrogatives. Here we restate each question and the contribution made by this paper: **1)** Do the particles and the sentences in the responses to negative declaratives and interrogatives bear particular intonational contours, if so which, and on which responses? The answer suggested by the data is that 74% of reverse responses bear the CC, and confirm responses almost always bear declarative intonation. **2)** Are some responses to negative declaratives and interrogatives more natural than others? The answer suggested by the data is all possible responses are judged equally natural (median 4 out of 5) with the exception of yes confirming responses (e.g. *yeah, I don't like movies*), which are judged slightly less natural (median 3 out of 5). **3)** Must polar particles be accompanied by sentences or sentence fragments for the responses to be acceptable? The present experiment does not answer this question, but suggests a future avenue of research: since polar particles in reverse responses usually bear the CC, and in confirm responses they almost always bear declarative intonation, it may be that speakers can distinguish the meaning of polar particles in response to negative declaratives and interrogatives on the basis of prosodic intonation. A perception study is required to test this.

We proposed a semantic characterization for the CC based on the distribution of the contour and on informal descriptions from Liberman and Sag (1974). The CC is modeled as a partial identity function that takes a proposition as an argument, and imposes the presupposition that the negation of that proposition is salient in the context. Future research will determine whether this characterization of the CC accurately captures the facts.

Acknowledgments

This paper started out as a class project by the first two authors in a class on laboratory linguistics taught in spring 2012 by the last author. Thanks to everyone in prosody.lab for their help in running this study, in particular to Erin Olson, Aron Hirsch and Elise McClay. For comments on an earlier draft of the paper, thanks to Bernhard Schwarz and Luis Alonso-Ovalle. For helpful discussion, thanks to Alanah McKillen, Brian Buccola, and Walter Pedersen, as well as the McGill linguistics syntax-semantics research group. This research was funded by grants from the FQRSC and from SSHRC, as well as from the Canada Research Chair Program.

References

R. Harald Baayen. 2008. *Analyzing Linguistic data. A Practical Introduction to Statistics Using R.* Cambridge University Press.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. In *Journal of Memory and Language* 68(3): 255-278.

Adrian Brasozeanu, Donka Farkas and Floris Roelofsen. 2011. Polarity particles and the anatomy of n-words. In *The proceedings of Sinn und Bedeutung* 16.

Noah Constant. 2012. English rise-fall-rise: a study in the semantics and pragmatics of intonation. *Linguistics and Philosophy*, 35:407–442.

Robin Cooper and Jonathan Ginzburg. 2011a. Negation in dialogue. In R. Arstein, M. Core, D. DeVault, K. Georgila, E. Kaiser, and A. Stent, eds, *Sem Dial 2011 (Los Angelogue): Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue 2011.*

Robin Cooper and Jonathan Ginzburg. 2011b. Negative inquisitiveness and alternatives-based negation. In M. Aloni, V. Kimmelman, F. Roelofsen, G. Weidmann-Sassoon, K. Schulz, and M. Westera, eds, *Logic, Language, and Meaning. Selected papers from the 18th Amsterdam Colloquium*, 32–41. Berlin Heidelberg. Springer.

Donka Farkas and Floris Roelofsen. 2012. Polar initiatives and polarity particle responses in an inquisitive discourse model. Ms, UC Santa Cruz, available via sites.google.com/site/inquisitivesemantics/papers-1/in-progress.

Anders Holmberg. 2012. On the syntax of yes and no in English. *Newcastle Working Papers in Linguistics* 18: 52-72.

Ruth Kramer and Kyle Rawlins. 2009. Polarity particles: an ellipsis account. In *The Proceedings of NELS* 39.

T. Florian Jaeger. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. In *Journal of Memory and Language* 59: 434-446.

Manfred Krifka. to appear. Negated polarity questions as denegations of assertions. In: Lee, Chungmin and Ferenc Kiefer, eds, *Contrastiveness and Scalar Implicature.* Berlin: Springer, available via amor.cms.hu-berlin.de/ h2816i3x/Publications/ NegatedPolarityQuestions.pdf

Manfred Krifka. 2013. Response particles as propositional anaphors. In *The proceedings of SALT* 23: 1-18.

D. Robert Ladd. 1980. *The structure of intonational meaning.* Bloomington, IN: Indiana University Press.

Mark Liberman and Ivan Sag. 1974. Prosodic form and discourse function. In *The proceedings of Chicago Linguistics Society (CLS)* 10: 402–415.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In Cohen, Philip R. and Jerry L. Morgan, eds, *Intentions in communication.* Cambridge, Mass.: MIT Press 271-311.

Michael Wagner, Lauren Mak and Elise McClay. 2013. Incomplete answers and the rise-fall-rise contour. In *Sem Dial 2013 (DialDam): Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue 2013.*

Effects of Belief and Memory on Strategic Negotiation

Markus Guhe
Alex Lascarides
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, Scotland
m.guhe@ed.ac.uk
alex@inf.ed.a.uk

Kevin O'Connor
Verena Rieser
School of Mathematical and Computer Sciences
Heriot Watt University
Edinburgh EH14 4AS, Scotland
kho30@hw.ac.uk
v.t.rieser@hw.ac.uk

Abstract

We present an empirical framework for testing negotiation strategies in a complex win–lose game that lacks any analytic solution. We explore how different belief and memory models affect trading and win rates. We show that cognitive limitations can be compensated for by being an 'optimistic' negotiator: make your desired trade offer, regardless of your beliefs about how opponents will react. In contrast, agents with good cognitive abilities can win with fewer but more effective offers. Corpus analysis shows human negotiators are somewhere in between, suggesting that they compensate for deficient memory and belief when necessary.

1 Introduction

Strategic negotiation is a type of non-cooperative conversation, which the Gricean view of cognitive agents fails to account for (Asher and Lascarides, 2013). In this paper we investigate negotiation dialogues as they occur during trading, within the example domain of the board game *The Settlers of Catan* (or *Settlers*, Teuber, 1995; see `www.catan.com`. We explore how human errors in beliefs, in particular forgetting, impact negotiating and trading behaviour, and take first steps towards building negotiation strategies that are effective in spite of deficiencies in beliefs.

Trading and bargaining is often modelled as rational actions between agents, all of whom maximise their *expected utilities*—an optimal trade-off between what they *prefer* (typically defined by a utility function) and what they *believe* they can achieve (typically defined via a dynamic Bayesian network; Savage, 1954). Solving a game problem involves finding *equilibrium strategies*: an optimal action for each player in that it maximises his expected utility, assuming that the other players perform their specified action (Shoham and Leyton-Brown, 2009). But this Savagean model of decision making is highly idealised and humans often deviate from it (Kahneman and Tversky, 1979; Ariely, 2008; Yong and Xinlin, 2012). Non-optimal human behaviour occurs in complex games, for instance, where existing algorithms for computing expected utilities can also break down. *Settlers* is one such game. It involves players conversing to negotiate trades over restricted resources. And even if an analytic solution for trading in *Settlers* were to exist, it doesn't necessarily match what humans do.

One response to this is to develop a symbolic model consisting of *heuristics* that match the strategies of expert human players (Thomas, 2003). But their effectiveness and correlation to human behaviour must be evaluated. Accordingly, we present here an empirical framework for devising and evaluating heuristics. We focus on testing agents with various cognitive limitations (e.g., memory loss), and we show that limited cognitive abilities can be compensated for by being an optimistic negotiator: make an offer for your most desired trade whatever your beliefs about its outcome. We then compare various computational agents with a corpus of humans playing *Settlers* (Afantenos et al., 2012).

2 The Settlers of Catan

Settlers is a win–lose board game for 2 to 4 players. Each player acquires resources (ore, wood, wheat, clay, sheep) and uses them in different combinations to build roads, settlements and cities on a board like the one shown in Figure 1. This earns points and the first player with 10 points wins. Players acquire resources in several ways, e.g., via the dice roll that starts each turn and through trading with other players—so players converse to negotiate trades. The dice rolls make

Figure 1: A game of *Settlers* in *JSettlers*.

future states non-deterministic, compelling players to assess the risk of their moves, including trading moves. A player's decisions about what resources to trade depends on what he wants to build; e.g., a road requires 1 clay and 1 wood. Trading decisions are also determined by estimates of what will most advance, or undermine, the opponents' strategies (Thomas, 2003). Players can also lose resources: e.g., a player who rolls a 7 can rob from another player. What's robbed is hidden from view, so players lack complete information about their opponents' resources. Because *Settlers* is a game of imperfect information, agents can, and frequently do, engage in 'futile' negotiations, which don't result in any trade. An agent that initiates a negotiation that doesn't result in a trade has in effect miscalculated the equilibrium strategies.

3 Related Work

There are several empirical approaches to modelling *Settlers*, but none of them includes trading or negotiation. Szita et al. (2010) and Roelofs (2012) use Monte Carlo Tree Search (MCTS), but on a simplified version of *Settlers* without any negotiation and trading between players. In contrast, our focus is on negotiation strategies and their interaction with cognitive ability.

Pfeiffer (2003) uses reinforcement learning to acquire *Settlers* strategies from game simulations. The results show that a mixture of hand-coded heuristics and learnt strategies challenges human players. This is encouraging, but it is unclear whether the heuristics are cognitively plausible or yield successful trading strategies. We address this here by evaluating different trading heuristics in a simulated game environment and comparing the results with what people do.

Related work on imperfect information games, such as poker and *Settlers*, emphasises the importance of modelling beliefs (Sweeney, 2012). An agent's beliefs about his opponents' intentions is known as *opponent modelling* or *nested beliefs* (Rieser et al., 2012). Vogel et al. (2013) have shown that nested beliefs are useful for reasoning about implicature-rich interpretations in dialogue. So the following experiments evaluate how the capacity to accurately model nested beliefs contributes to negotiating, trading and winning.

4 Planning in *JSettlers*

We use an open source implementation called *JSettlers* (`jsettlers2.sourceforge.net`, Thomas, 2003). *JSettlers* is a client–server system: a server maintains the game state and passes messages between each of the players' clients, which can run on different computers. Clients can be human players or computer agents. Here, we report on simulations between computer agents.

The *JSettlers* agent, which we call the *original* agent, goes through multiple phases after the dice roll that starts its turn:

1. Deal with game events: e.g. placing the robber; acquiring or discarding resources.
2. Determine legal and potential places to build.
3. Estimate the time required to build pieces on legal places (the ETB).
4. Compute the *Best Build Plan* (BBP): a sequence of build actions that achieves 10 points in the shortest estimated time (ignoring how opponents might hinder your plans).
5. Try to execute the BBP, including negotiating and trading with other players and/or the bank or a port.

As we are exploring how various cognitive limitations impact decisions to negotiate and how that affects trading and winning, all our agents adopt the same *build* strategy: i.e., steps 1–4 remain constant, while step 5 differs across agents. We first describe these common steps.

Agents sort resources into *needed* vs. *not needed* given their BBP. When considering whether to offer an unneeded resource for a needed one (and likewise when considering whether to accept an offer addressed to them), an agent compares the *Estimated Build Time* (ETB) of the offer against that of its *Best Alternative to Negotiated Agreement* (BATNA), the latter being no trade at all, a bank trade (at a 4:1 ratio) or a port trade (at a 3:1 or 2:1 ratio, depending on the port). All our agents make the offer only if the former ETB is shorter than the latter.

Sufficient conditions for making an offer differ among our agents, however. By using the same ETBs and build policies, all our agents have the same 'instrinsic' trading preferences, but differ in extrinsic preferences because of different beliefs about the offer's outcome. Their differing beliefs stem from their different *cognitive capacities* (how much evidence they have for predicting an offer's outcome), and their different ways of handling missing evidence (see Section 5 for details).

Agents have three possible responses to a trade offer during a negotiation: to *accept* it (and enact the trade), to *counteroffer* (which may be a completion of a partial offer) or to *reject* it. An agent accepts an offer if the offer is executable and its ETB is less than that of its BATNA; he counteroffers if the offer's ETB is not less than that of its BATNA but there is an offer that he hasn't already made that satisfies the agent's sufficient conditions for offering; otherwise he rejects the offer.

5 Negotiation Strategies

To win, an agent needs an effective *trading* strategy and an effective *negotiation* strategy. An effective trading strategy is one that increases the player's likelihood of winning: i.e., on average, his choices of when and what to trade help him more than they hurt him. An effective negotiation strategy increases the likelihood that negotiations culminate in an effective trade. Our initial experiments demonstrate that the implemented *JSettlers* trading strategy is effective (see Sections 6.2 and 6.3). So in order to evaluate distinct negotiation policies and belief models with appropriate controls, we make all agents adopt the same effective *JSettlers* trading choices.

A trade offer has a non-deterministic outcome because it depends on the opponents reaction: a desired outcome is that the offer is accepted (so the trade is enacted) but without that trade helping the opponent more than it helps the proposer. Here, this means that the opponent has fewer than 8 points and his BBP doesnt block the proposers own BBP. Thus agents should use evidence, both past and present, to estimate the opponents (hidden) resources and BBP, to infer whether an offer will have this desired outcome.

In this paper, we manipulate the cognitive capacity of an agent via how much evidence he has for predicting an outcome; and we manipulate how optimistic or pessimistic he is about an outcome when evidence is inconclusive. So overall, we manipulate what evidence an agent uses, and how he uses it. In total we investigate 10 conditions, see table 1.

On **cognitive capacity**, we investigate at one extreme agents who are *omniscient* about the opponents' resources and/or BBPs (making the game state fully observable), and at the other extreme agents who lack any evidence at all (either past or present) for inferring them. Within these two extremes, we implement agents who use past and present evidence to estimate the opponents' resources and BBPs, but they *forget* past evidence after a certain time. We give the details in Section 6.

We then distinguish three ways of using evidence to predict the outcomes of dialogue moves. An **optimistic proposer** makes a trade offer with the best possible ETB regardless of the evidence for its outcome; thus, he in effect ignores evidence that's against the desired outcome even when it exists!

Alternatively one can be less optimistic, making a trade offer only if the available evidence yields a belief in its desired outcome. In fact, we implement two non-optimistic agents, which handle *missing evidence* differently. The original *JSettlers* agent is what we call a **cautious proposer**: he makes an offer only if he believes it will have the desired outcome, but he defaults to this belief when the evidence is inconclusive. That is, when he has insufficient evidence to infer certain information about his opponents' resources and BBPs, he simply assumes favourable values.

This contrasts with a **pessimistic proposer** who makes a trade offer only if he believes it will have the desired outcome, but unlike the cautious proposer he defaults to believing it won't in the absence of information to the contrary. Thus, by default the pessimist assumes the trading partner does not have his desired resource or is not willing to sell it or does not need what the pessimist is offering.

Finally, we implement agents who issue *partial* offers like (1a), as opposed to only making *complete* offers like (1b):

(1) a. I need clay.
b. Max, will you give me 1 clay for 1 wood?

The original *JSettlers* agent only makes *complete* offers like (1b), but this isn't human-like (see Section 6.8 for details): only 4.7% of the offers in the corpus we collected of people playing *Settlers* are as specific as (1b); 23.1% specify the resources but not their quantity; and the most frequent type of offer, at 34.8%, is a partial offer specifying the receivable resource, as in (1a).

In addition to partial offers, we evaluate different proposer-types (optimistic, cautious, pessimistic) with different cognitive abilities (omniscient, ignorant, forgetful) in terms of their effectiveness as negotiators and their chances of winning. We started with Thomas' (2003) **original** *JSettlers* agent. We modified the code slightly, e.g., by improving the initial placement of pieces and fixing a number of bugs that, for example, affected the automated running of large numbers of games in our simulations. However, the main planning and trading mechanisms are unchanged, and this agent remains a cautious proposer.

This original agent is neither omniscient nor ignorant, but he's not human-like: He has a perfect memory, never forgetting past evidence, and so maintains a perfect model of the opponents' resources until a 7 is rolled. At this point, the robbed player loses a resource to the robbing player, and any player with 7 or more resources must discard half of them. Which resources a player loses is unobservable to agents not involved in the transfer, and so the *JSettlers* agent downdates his beliefs for *all* resources of the affected player to *unknown*. This extreme form of belief change is also not human-like, as humans would still keep an hypothesis of potential resources owned by this player. Here, we investigate how human memory, in particular forgetting, influences the effectiveness of various distinct negotiation strategies, showing that some of these penalise a player with deficient beliefs while others do not. We thus make the first steps towards building negotiation strategies that are effective in spite of human-like errors in beliefs. In future work, we will also investigate how more elaborate forms of belief update and revision (after unobservable events) will influence negotiation strategies.

The original agent is not strong but is at least in the ballpark of human performance. Thomas (2003) performed an evaluation where in each game three agents (agents without our bug fixes and improvements) played one human player. The human player won about 50% of the games (and each agent about 17%).

6 Experiments with Modified Agents

6.1 Method

Simulations for testing a particular belief model and strategy for proposing trades all consist of 1 modified agent playing 3 original agents in 10,000 games. So the null hypothesis is that each agent wins 25% of these 10,000 games. To carry out these simulations, we created a simulation environment for *JSettlers*. The server and the 4 agents all ran on the same machine, and a simulation of 10,000 games took about 0.5–1h on a desktop computer.

In addition to measuring the win rate, we analyse the agents' negotiating and trading behaviour: the number of offers they made, the total number of successful offers (i.e., how many trade offers resulted in a trade), and the total number of trades with other players, i.e. including trades resulting from accepting other players' offers. Finally, the proportion of an agent's offers that are successful gives a rough measure of how accurately he estimates an equilibrium trading move. Table 1 gives an overview of the results.

Due to the large number of games in each simulation even small differences can be significant (as long as the standard deviation is also reduced). At the same time, in the simulations reported here there are no significant differences between the three instances of the original agent, i.e. all differences result from agent modifications. For the simulations, we test significance of win rates against the null hypothesis (25%) by using the z-test; we analyse differences in trading behaviour among opponents via paired t-tests for all combinations of opponents (in fact, there were no significant differences for the offering/trading measures between any two original opponents); and we use a significance threshold of $p < 0.01$. We now describe the simulations in detail, in the order in which they are given in Table 1.

	modified agent					averages for the 3 original agents				
	win rate	offers	success-ful offers	total trades	successful offers/ offers	win rate	offers	success-ful offers	total trades	successful offers/ offers
random BBP	**0.022**	**21.9**	**5.1**	**9.0**	**0.23**	0.326	14.2	3.9	8.1	0.27
non trading	**0.127**	**0.0**	**0.0**	**0.0**	---	0.291	17.7	3.1	6.2	0.18
omniscient · resource	0.244	**7.5**	3.7	7.5	**0.50**	0.252	13.6	3.8	7.6	0.28
omniscient · BBP	0.243	13.8	**3.6**	7.6	**0.26**	0.252	13.7	3.8	7.6	0.28
omniscient · BBP/resource	0.243	**7.3**	**3.5**	7.5	**0.48**	0.252	13.8	3.9	7.5	0.28
ignorant · cautious	0.241	**28.5**	**3.6**	7.7	**0.13**	0.253	13.9	3.9	7.7	0.28
ignorant · pessimist	**0.162**	**0.0**	**0.0**	**4.2**	---	0.279	14.7	4.2	7.0	0.28
ignorant · optimist	0.255	**27.1**	**3.7**	**7.8**	**0.14**	0.249	13.8	3.9	7.6	0.28
ignorant · caut. 65% cap	**0.203**	**12.7**	**1.9**	**6.1**	**0.15**	0.266	14.2	4.0	7.3	0.29
partialising initially	0.240	**19.4**	**2.7**	**7.9**	**0.14**	0.253	14.7	4.2	7.7	0.29

Table 1: Results of 1 modified agent playing 3 original agents. Results for the modified agents are shown in the left half; the corresponding averages of the results for the 3 original agents are shown in the right half. Significant results ($p < 0.01$) are in bold.

6.2 Quality of Build Strategy

An agent's trading and negotiation strategies are based on a best build plan (see step 4 in the agent's planning procedure, Section 4). In order to establish whether we can retain the original agent's build plan for our experiments, we evaluate its quality by testing it against a **random BBP agent**, who chooses his 'best' build plan randomly.

Results and Discussion. This agent wins only 2.2% of the games, even though he makes more trade offers, more successful offers and more trades (using the same negotiation and trade strategy as the original agent). So the original agent's build strategy is an improvement over a random baseline. Together with the evaluation in Thomas (2003), we can therefore assume that all of our agents, which retain the *JSettlers* build strategy, have decent build plans.

6.3 Benefits of Trading

To establish that trading contributes to winning, we created the **non-trading agent**, which is like the original agent except that he never trades with other players but only with the bank or a port.

Results and Discussion. The non-trading agent wins only half as many games as his opponents, providing strong evidence that the *JSettlers* trading policy is effective and contributes to winning. (The agent makes more trades with the bank and ports (9.6) than his original opponents (4.9), but this does not compensate for not trading with agents.) So the agent's *preferences* over possible trades, defined by his ETBs and BBP, correlates with his chances to win. Since we never change these calculations, any changes to win rates will stem from how effective the *negotiation strategy* is in achieving a trade.

6.4 Beliefs: Omniscience

To explore how useful accurate beliefs about the opponents' resources are, we tested a **resource omniscient agent**, who *directly observes* his opponents' resources but remains a **cautious proposer**. So he never has defeasible beliefs about resources, but may default to a belief that his opponent has a favourable BBP. We make the resources observable by getting the (original) opponents to declare them at the start of each turn (note that we don't allow deception in our simulations; the role of deception in *Settlers* is future work). Moves that declare resources, or lack of them, are attested frequently in the human *Settlers* corpus (Afantenos et al., 2012), generally via responses to questions about offers; e.g., *I've got clay* in response to *What will you give me?*, or *I don't have any* in response to *I need wood.*

We also tested a **BBP-omniscient agent**: again a **cautious proposer** but one for whom his opponents' BBPs are always observable. Again, these are observable because agents declare them (*I intend to build a road.*). These are attested but rare in the *Settlers* corpus.

Finally, we test an agent that is **omniscient on both BBPs and resources**.

Results and Discussion. None of these three agents have significantly different win rates than their original opponents. However, they all have a more effective negotiating strategy, with a higher proportion of their offers being successful. But the resource-omniscient agents make far fewer offers overall, and the BBP-omniscient agent makes slightly but significantly fewer successful trade offers.

With respect to human games, entering fewer 'futile' negotiations is relevant because human players can easily get annoyed when players make many trade offers, and in particular offers where it is obvious that it won't be accepted.

It may seem counterintuitive that making hidden parts of the *Settlers* game state observable fails to improve the win rate. We believe that this happens for three reasons. First, the original agent's perfect memory gives him good resource-tracking capabilities: the only relative advantage of the resource-omniscient agent comes after a 7 has been rolled.

Secondly, and perhaps more importantly, all of our agents, both modified and original, are only willing to negotiate for their best possible trade, and do not consider entering a negotiation for a 'second best' possibility should they believe that their best possible trade isn't achievable. This aspect of the negotiation strategy hurts the omniscient agent: since he never defaults to favourable values, he starts a negotiation less frequently than an original agent would (in our simulations, around half the time). This denies the omniscient agents the chance to consider *counteroffers* that may not be best but are nevertheless effective and so acceptable. But the non-omniscient agent gets relatively more opportunities to consider such counteroffers. In future work, we plan to investigate how adapting the negotiation strategy to allow initiating a negotiation for a 'second best' build plan would enhance the win rates.

Finally, making declarations of BBPs seems to be largely redundant: all agents use the same evaluation function for computing everyone's BBPs, and for the opponents' BBPs this function draws only on the observable part of the game state and beliefs about the opponents' (hidden) resources, which even for the original agent are relatively accurate. This redundancy also explains why the **BBP/resource-omniscient** agent performs to much the same level as the resource omniscient agent. In future work, we plan to investigate how declaring build plans when agents have distinct build policies—as humans players invariably do—impacts game performance.

6.5 Beliefs: Ignorant

The **ignorant agent** does not track the other agents' resources at all. He can deal with this lack of knowledge in 3 ways: being **optimistic** (assume all opponents have all resources), being **cautious** (treat all opponents' resources as unknown but track the overall amount, and default to assuming they have the desired resource if this doesn't conflict with knowledge of how many resources they have) or being **pessimistic** (assume all opponents have no resources). So the pessimistic agent never makes an initial offer (because he believes that the offer will be rejected) but may accept offers from others. In contrast, the optimistic and cautious agents make many offers.

Results and Discussion. The ignorant pessimist wins significantly fewer games and makes fewer trades (and no trade offers, by design). In contrast, the cautious and optimistic proposers can compensate for the deficient belief model by making twice as many trade offers. Thus, while it pays to have a relatively accurate belief model and to reason about likely outcomes of offers, there are also negotiation strategies that compensate for a deficient belief model, which involve *ignoring* the risk of your desired offer having an undesired outcome, i.e. not being accepted.

Capping the number of trade offers. Because the optimistic and cautious proposers compensate for ignorance by making many trade offers, we tested whether the driving factor is the quantity or the quality of the offers, by capping the number of offers the cautious ignorant agent can make to the same level as his original opponents. We implemented this by letting the ignorant agent make the decision on whether to make a trade offer but then only 'allowed' it to actually go through with making it 65% of the time. As can be seen in Table 1, this agent fares much worse than the equivalent 'non-capped' ignorant agent, showing that he is able to approach the winning rates of his less belief-deficient opponents only by counteracting the lower quality in his trade offers by increasing their quantity. Note also that in this simulation, the original opponents make fewer trades because the ignorant capped agent makes fewer acceptable offers.

6.6 Negotiation Strategy: Partial Offers

As mentioned before, a major difference between the way the *JSettlers* agents and human players negotiate is that people often make partial offers like example (1a). People typically do not make a complete trade offer immediately, but incrementally negotiate towards a more specific, and mutually acceptable, offer.

For this reason, we tested a **partialising agent** that initially computes a complete trade offer but then partialises it to only specify the resources it wants—i.e., it starts a negotiation with (1a), as opposed to (1b). It then reverts back to the original negotiation strategy. In this way, we test how our agents fare when adopting a human-like initial offering strategy.

Results and Discussion. Making an initial partial offer does not affect the agent's chances of winning and results in a small but significant increase in the number of trades it makes. But this agent does make many more trade offers than the original agents of which fewer are successful. The latter is due to the fact that agents cannot accept partial offers: typically, the complete offer that complements the initial partial offer was made by one of the original agents.

6.7 Memory: Forgetting Beliefs

The original agent and our modified agents so far do not have a realistic model of human memory: they do not forget any observed information. Therefore, we made the agent forget his beliefs about his opponents' resources after a certain time. For example, if the agent did not receive any new evidence about player-2's clay resources, the belief was set to 0 for the **forgetting pessimist proposer** and to 2 for the **forgetting cautious proposer** (so the cautious proposer assumes a favourable value in the absence of evidence to the contrary). In future work we will replace this by a more realistic and cognitively adequate memory model.

Because the *JSettlers* system does not maintain a global time, we used the number of messages sent by the *JSettlers* game server to approximate passing time. In a typical game, the server sends between 3500 and 5500 messages. We varied the time after which an agent forgets a belief from 5 to 1000. Note that ignorant agents are agents with a forgetting latency of 0.

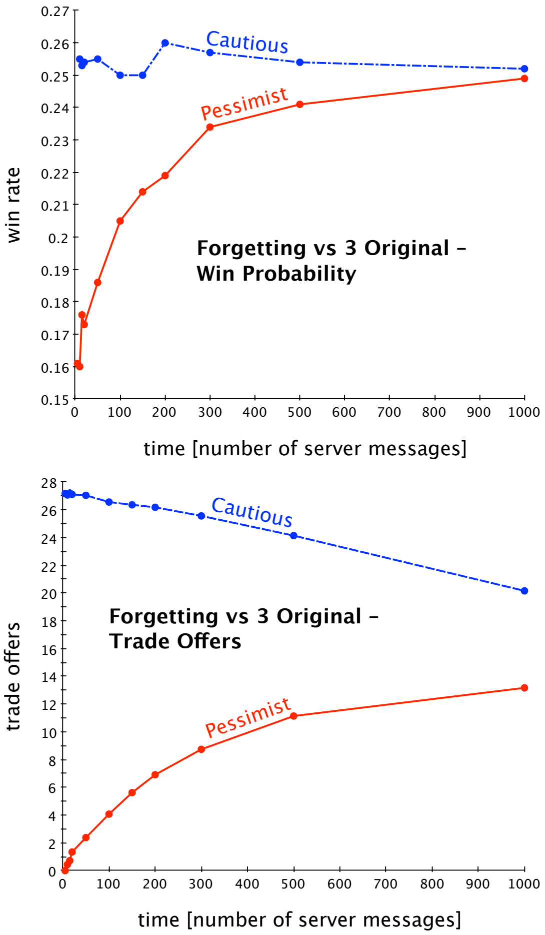


Figure 2: Win rates and number of offers over the latency before information is forgotten.

Results and Discussion. Figure 2 shows how the win rate and the number of trade offers changes with the time an agent can remember his beliefs. The cautious forgetting agent can, once more, compensate his lack of knowledge by assuming a desired outcome will occur in the absence of evidence to the contrary. And while the number of successful offers and total trades is constant for the cautious proposer and is the same as that of the original agent, the more forgetful he is, the more he makes unsuccessful trades offers and his negotiations become less effective.

The pessimistic forgetting agent makes fewer trade offers and wins fewer games the more forgetful he is. Only with a considerable ability to remember information (i.e. only if he forgets information after about 500 to 1000 time steps, which is already a considerable part of the entire game) does his performance approach that of the original agent.

Thus, only if a forgetful agent adopts favourable

	offers	successful offers	total trades	TIO	trades matching TIO	trades not matching TIO	successful offers/ offers
original	13.5	3.8	7.6	6.5	2.5	1.1	0.28
ignorant · cautious	30.6	3.8	7.6	10.2	1.8	1.9	0.12
omniscient · BP/resource	13.7	3.8	7.7	6.5	2.6	1.1	0.28
partialising initially	22.8	3.8	7.5	6.6	2.6	3.6	0.17
forgetting pessimist 200	6.7	2.8	5.6	4.0	2.1	0.6	0.42
Novice Corpus	*21.5*	*2.6*	*5.2*	*10.4*	*1.9*	*1.4*	*0.12*

Table 2: Results of 4 agents of the same type, and the corpus of novice players. TIO are turn-initial offers.

beliefs in the absence of actual information, can he compensate for his deficient memory.

6.8 A Comparison with Human Data

We now compare the simulated behaviour against the negotiating and trading behaviour of people playing *Settlers*. Our human data, shown in the last row of Table 2, is taken from an annotated corpus of humans playing *Settlers*, where detailed information about bargaining moves (offers, counteroffers, acceptance, rejection, etc.) and associated information about giveable and receivable resources that offers express are recorded (Afantenos et al., 2012).

As a first step, we used the annotated data from 7 games, where all players were new to the game, i.e. had not played *Settlers* before participating in the study. In future work, we will also compare the performance against players with varying degrees of expertise. As all players in the current set of games are members of a homogeneous population (novice *Settlers* players) we ran simulations where 4 agents *of the same type* play against each other. We only did this for the agents that are most interesting for a comparison (see Table 2).

Results and Discussion. The results from the simulations and the corpus given in Table 2 show that there is not one single agent from those we have investigated so far that exhibits the same behaviour as the novice players on all factors. The number of human trade offers is closest to the agent that initially partialises his trade offers; the successful trade offers and total trades are similar to the forgetting pessimist (with a forgetting latency of 200); and looking at the turn-initial offers (i.e. considering only the first offer of any given turn) and the ratio of successful offers over offers, the novices resemble the ignorant cautious (or the very similar ignorant optimistic) agent.

Thus, although novice players make very many trade offers, they are at the lower end of the spectrum when it comes to making successful offers, which in turn limits the number of trades they make. Note that the high number of trade offers is not simply a result of adopting a negotiation strategy of making partial offers (which, by definition, require at least one other offer to complete it): the high number of turn-initial offers shows that the number of game turns in which people try to trade is at the upper end of the spectrum.

These results are consistent with our previous explanations. People don't have a perfect memory (and for novice players, who have to keep track of many unfamiliar aspects of the game, this seems even more true). So they forget information. On the other hand they make many more attempts to trade, which is consistent with our suggestion that one strategy to compensate for an imperfect memory (up to a point) is to negotiate more.

7 Conclusions and Future Work

Strategic conversation, where the agents' goals diverge, is typically modelled as a game with a known analytic solution, where standard algorithms for identifying optimal actions apply. But human behaviour often diverges from game theoretic solutions and furthermore such models do not apply to noncooperative dialogues in the context of a complex game like *Settlers* that lacks any analytic solution. We, therefore, presented an alternative approach—an empirical framework in which on can evaluate how distinct dialogue strategies fare in the fact of distinct belief and memory models, including models that exhibit human-like errors like forgetting.

We first established that the existing *JSettlers* trading strategy correlates with winning. We then experimented with various models of the means for achieving such trades, and compared these models with the behaviour exhibited in a corpus of people playing *Settlers*.

Our agents varied on the extent to which they have accurate *beliefs* about hidden aspects of the game state, and the extent to which they can *remember evidence* for those hidden values. Agents with limited accuracy in their beliefs have three coping strategies for handling missing evidence: being *optimistic* and always assuming that the desired offer will have the desired outcome; being *cautious* (e.g., the original *JSettlers* agent) by using the available evidence to estimate hidden values but defaulting to favourable hidden values when evidence is inconclusive; and being *pessimistic* by assuming unfavourable values when evidence is inconclusive.

Belief and memory models affect the number and the quality of trade offers that agents make. An agent with perfect knowledge of the other agents' resources and a non-optimistic negotiation strategy makes offers that are much more likely to result in trades, but he fails to outperform his opponents' win rate, even though the opponents' belief models are more fallible. We speculate that this is because the opponents are more likely to enter into a negotiation (given that they default to assuming it will have the desired effect), and so increases the likelihood that they make an advantageous trade, if not the initially desired trade. Thus some ignorance can be bliss!

Indeed, being completely ignorant can be compensated for by being optimistic. This results in a much less efficient negotiator but more trades overall. On the other hand, capping the number of offers inhibits the advantages of this strategy. We can draw similar lessons from the forgetting agents, which show that (all else being equal) being able to remember more about your opponents' resources increases your chances to win. Or just be more optimistic (and less efficient).

We also took our first steps towards comparing our computational negotiation models with the strategies deployed by people playing *Settlers*. We showed that the predominant way of making partial offers observed in the human corpus does not change the win rate of our existing agents, and the observed changes in the number of offers and successful trade offers are consistent with our account. We also showed that while none of our agents directly models human performance, their performance is comparable in many ways. Of our agents, the ignorant optimistic agent is closest to the novice human performance exhibited in the corpus. In future work we will create cognitively more plausible agents and evaluate them by letting them play against humans.

While it was necessary for all agents to use the same build strategy (except for the random agent) so as to reduce the number of variables in our simulations, this is clearly an oversimplification in that the agents enjoy almost perfect predictions of their opponents' build plans.

In future work, we will explore the relative merits of revealing vs. concealing information about intentions when the agents all deploy distinct build strategies. We will also enhance the negotiation strategies by allowing agents to initiate a negotiation for a 'second best' trade when they believe their best trade won't have the desired effects, and we will investigate the benefits and costs of deception in a trade negotiation.

Acknowledgments: This work is supported by ERC grant 269427 (STAC). Thanks to Robin Hill for discussions on statistics and to Nicholas Asher, Oliver Lemon and 3 anonymous reviewers for comments and feedback.

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. Modelling strategic conversation: model, annotation design and corpus. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (Seinedial)*, Paris, 2012.

D. Ariely. *Predictably Irrational: The Hidden Forces That Shape Our Decisions.* Harper-Collins, 2008.

Nicholas Asher and Alex Lascarides. Strategic conversation. *Sematics & Pragmatics*, 6(2):1–62, 2013.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

M. Pfeiffer. Machine learning applications in computer games. Master's thesis, Technical University of Graz, 2003.

V. Rieser, O. Lemon, and S. Keizer. Opponent modelling for optimising strategic dialogue. In *16th Workshop on the Semantics and Pragmatics of Dialogue (SeineDial'12)*, 2012.

G. Roelofs. Monte carlo tree search in a modern board game framework. research paper available at `umimaas.nl`, 2012.

L Savage. *The Foundations of Statistics*. John Wiley, 1954.

Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic and Logical Foundations*. Cambridge University Press, 2009.

N. Sweeney. *Applying Machine Learning Techniques to an Imperfect Information Game*. PhD thesis, School of Computing, Dublin City University, 2012.

I. Szita, G. Chaslot, and P. Spronck. Monte-carlo tree search in Settlers of Catan. In H. van den Herik and P. Spronck, editors, *Advances in Computer Games*, pages 21–32. Springer, 2010.

K. Teuber. *Die Siedler von Catan: Regelheft*. Kosmos Verlag, Stuttgart, Germany, 1995.

R. Thomas. *Real-time Decision Making for Adversarial Environments Using a Plan-based Heuristic*. PhD thesis, Department of Computer Science, Northwestern University, 2003.

Adam Vogel, Christopher Potts, and Dan Jurafsky. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Z. Yong and W. Xinlin. Intrinsic preferences, revealed preferences and bounded rational decisions. *Systems Research and Behavioural Science*, 2012. doi: 10.1002/sres.2120.

Modelling Expectation in the Self-Repair Processing of Annotat-, um, Listeners

Julian Hough and Matthew Purver
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
`{j.hough,m.purver}@qmul.ac.uk`

Abstract

This paper describes a statistical corpus study of self-repairs in the disfluency-annotated Switchboard corpus which examines the time-linear nature of self-repair processing for annotators and listeners in dialogue. The study suggests a strictly local detection and processing mechanism for self-repairs is sufficient, an advantage currently not used effectively under the bonnet of state-of-the-art automatic disfluency processing. We then show how simple local fluency measures using modified language models can be strongly indicative of repair onset detection, and how simple information theoretic measures could characterize different classes of repairs.

1 Introduction

Statistical language modelling for self-repair has enjoyed good results for accurately detecting edited words within repairs (Heeman and Allen, 1999; Charniak and Johnson, 2001; Johnson and Charniak, 2004; Georgila, 2009; Zwarts et al., 2010; Qian and Liu, 2013). However, these successful systems ignore the classification of the repair's function and interpretation; furthermore the models used are generally computationally complex, over-predictive, and unrepresentative of a listener's incremental interpretation process, raising questions of psychological plausibility.

Beginning with classification, we consider the structure and taxonomy of first-position self-repairs, following the annotation scheme first proposed by Shriberg (1994) and the Switchboard disfluency corpus (Meteer and Taylor, 1995) annotation protocol:

$$\underbrace{\text{John and Bill}}_{\text{original utterance}}\ \ \underbrace{\text{[like}}_{\text{reparandum}}\ \ +\ \ \underbrace{\{\text{uh}\}}_{\text{interregnum}}\ \ \underbrace{\text{love]}}_{\text{repair}}\ \ \underbrace{\text{Mary}}_{\text{continuation}}$$

In addition to this structural vocabulary, from here on we consider the *repair onset* to be the first word after the (possibly null) interregnum, and the *interruption point* as the transition labelled '+ ' between the reparandum and the repair. Within this schema it is possible to distinguish three main classes of repair:

(1) "But one of [the, + the] two things that I'm really…"
Repeat (sw4356)[1]

(2) "Our situation is just [a little bit, + kind of the opposite] of that"
Substitution (sw4103)

(3) "…the bank was suing them [for, + { uh,}] because they went to get…"
Delete (sw4356)

Intuitively, a repair seems likely to be interpreted as a delete (3) if the following word (the repair onset) has no substitutional relation with its reparandum before the interruption point, having an overriding or cancelling effect; substitutions (2), in contrast, do exhibit some substitutive property or parallelism; and verbatim repeats (1), almost trivially, exhibit complete parallelism.

The interpretation of self-repairs by both annotators assigning bracketing on transcripts, and listeners assigning an interpretation function during dialogue, is not trivial. One could argue that simply checking for verbatim repetition for repeats, syntactic constituent identity for substitutions – see Levelt (1983) – and otherwise positing a delete, is a sufficient classification protocol. However there are many different possible subclasses, and gradient effects are exhibited in judgements of the classification of delete or substitution. As we have found in preliminary annotation experiments, annotators do not agree on each decision. The following examples show possible alternative interpretations (italicized) to the Switchboard annotations:

---

[1] sw* are conversation numbers in Switchboard.

(4) “and [there’s, + *?]* it’s] completely generic.”
Substitution or delete? (sw4619)

(5) “a matter where priorities are [at, +] placed.*?]*”
Delete or substitution? (sw4360)

In terms of the incremental dialogue semantics of these different forms, as Ginzburg et al. (2013) discuss, there is a broad difference between *forward-looking* (verbatim repeats, filled pauses/editing terms) and *backward-looking* interpretations (reformulation such as (2)). Deletes, utterance-initial forms of which are often called *restarts*, are more destructive than substitutions as they are driven by continuing the original utterance, rather than replacing or modifying the reparandum. A speaker or annotator may infer a more dramatic change of content in processing a delete. Also in on-line detection, as they do not adhere to well-formedness rules (Levelt, 1983), one must use other mechanisms to process them.

We maintain the classification distinction between substitutions and deletes, but the need for gradient judgements between these classes is clear due to the possible different interpretations of (4)-(5). Some repairs are more prototypical of their class than others. We address this in section 5.

The second issue we wish to address is the time-linear way in which people process repairs, a constraint which rule-based disfluency detection models do not prioritise – even if they are embedded in incremental systems – as we will discuss below. In consideration of working memory constraints, it is much more likely that repair operations begin once the repair onset is detected, rather than constantly predicting a reparandum before any disfluency has been encountered. Resolution can still be as fast and automatic as psycholinguistic evidence suggests (Brennan and Schober, 2001), but without maintaining all possible repair paths.

We investigate the intuitions of a local self-repair detection and resolution mechanism with gradient interpretation through a corpus study and language modelling. Our corpus study in section 4 observes the frequency of the three main classes and their subtypes in Switchboard, and the interactions of repair class distribution with features of their local utterance contexts. We then present a potential model of incremental repair detection and and interpretation in section 5, based on information theoretic measures.

2 Previous Work

Corpus analysis of Switchboard Shriberg and colleagues (1994; 1996; 1998) have done extensive work annotating and analysing the Switchboard corpus for repairs, editing terms and filled pauses, using a reliable disfluency annotation scheme (Meteer and Taylor, 1995) (see above). Shriberg (1994) creates a taxonomy of disfluency types: filled pause (FP), articulation disfluency (ART), substitution (SUB), insertion (INS), deletion (DEL), repetition (REP), hybrid disfluency (HYB) and conjunction (CON), the last of which occurs between speaker utterances. HYB is an important member of Shriberg’s taxonomy due to the plethora of combinations of these repair operations in Switchboard, as we will show below.

Shriberg (1996) compares the distribution of disfluency types across three different dialogue domains including Switchboard. The most common type in all three domains is FP followed by REP. DEL, defined as a repair containing at least one deleted word with no insertions or substitutions, and SUB, defined as a repair having at least one substitutive relation to the reparandum with no deletes or insertions, were ranked 3rd and 4th in Switchboard respectively.

In terms of incremental processing, Shriberg showed an interaction between the position of the interruption point and the disfluency type: per-word rates by position showed that the three most common disfluencies (FP, REP, and DEL) were much more likely to occur in initial position than in medial position. The remaining types appear to be roughly equally likely in initial and medial positions. Furthermore Shriberg and Stolcke (1998) investigate retraces, which are either verbatim repeats or repairs with one or more repeated words. Fitting parameters over the entire disfluency-tagged corpus, there is a logarithmic decay in the likelihood of retracing back one more word as the number of words since the last utterance or repair boundary increases. Speakers rarely retrace more than one or two words. This relationship supports a claim for a very local strategy for repair resolution.

Statistical self-repair detection In state-of-the-art self-repair detection on transcripts, Qian and Liu (2013) achieve the best reported performance on the Switchboard disfluency test corpus, achieving an f-score for detecting reparandum words of 0.841. They use a three step detection system using weighted Max-Margin Markov (M^3) networks: (1) detection of edit-terms/fillers/interregna (2) detection of reparandum words, and (3) refining the previous steps, using a cost-sensitive error function. Georgila (2009) introduces a post-processing method of Integer Linear Programming (ILP) to improve overall accuracy of various off-the-shelf methods, reporting an f-score for detecting reparandum onset words at 0.808 and repair onsets at 0.825 for a CRF model. While these results are impressive, the systems do not operate incrementally: they maximise the overall likelihood of tag sequences in utterances, using utterance-global constraints, rather than focussing on incremental accuracy.

Zwarts et al. (2010) describe an incremental version of Johnson and Charniak (2004)'s noisy channel model. The detector uses a bigram language model trained on roughly 100K utterances of reparandum-excised Switchboard data for its "cleaned" language model. Its channel model is a statistically-trained S-TAG which has simple reparandum-repair alignment rules for its non-terminals (copy,delete,insert,substitute), parsing all possible repair structures for a given utterance hypothesised in a chart, before pruning the unlikely ones. It performs equally well as the non-incremental model by the end of each utterance, achieving an f-score of 0.778 for the Switchboard disfluency task, and is modified to make detections early. They report the novel incremental evaluation method of *time-to-detection* for correctly identified repairs, achieving an average of 7.5 words from the start of the reparandum and 4.6 from the start of the repair phase, longer than the average repair length. They also introduce *delayed accuracy*, a word-by-word recall evaluation of the gold-standard disfluency tags from the point reached utterance so far, reporting recall in one word histories being 0.578, steadily increasing word-by-word until 6 words back where it reaches 0.770.

An earlier incremental system was Heeman and Allen (1999)'s multi-knowledge source approach which employs templates of repair structures within a complex incremental language model. This performs slightly worse than the noisy-channel approach above at detecting reparandum words (recall 65.9% and precision 74.3%), with the sparseness of the data providing problems for templates– the fact that the repeat sequence, w_1, w_1 is the most common repair structure may be very useful for an incremental classifier, but there is a long tail in the distribution of repair structures: they report that 1,302 modification repairs (non-deletes) take on 160 different repair structures in the TRAINS corpus, with only 47 (29.4%) occurring at least twice. To combat this they use over-prediction of templates, initially providing high recall with low precision, then filter out unlikely candidate repair structures using lexical, POS and intonation features. They include a feature encoding that a repair has already been detected in the utterance: in TRAINS, 35.6% of repairs overlap. Utterance-initial cancelling repairs (re-starts), were particularly problematic to identify – we suspect through lack of POS- or word-level parallelism and available templates, which can be exploited in repeats and substitutions, but not for deletes. Heeman and Allen also report very high accuracy for detecting discourse markers/editing terms (both as interregna and as forward-looking repairs), identifying 97% of them with 96% precision.

3 Approach: locally triggered repair detection and classification

In the popular automatic detection task, while incremental systems exist, they use over-prediction, large chart storage and filtering (Zwarts et al., 2010; Heeman and Allen, 1999). A parsing chart used solely for disfluency structures positing every possible repair path grows approximately cubically with the length of the utterance. Also, (Zwarts et al., 2010)'s TAG parser also has a run-time complexity of $O(N^5)$. This complexity blow-up seems cognitively implausible, particularly given the relative sparsity of repairs. In addition, these approaches cannot easily deal with processing embedded repairs realistically, as a stack of charts would be required, further increasing complexity– consequently these are ignored in training (Johnson and Charniak, 2004). Rather than positing all possible repair alignments, intuitively, a listener is almost certain an utterance is a non-repair before the repair onset, so a backtracking mechanism employed upon interruption point detection seems more plausible. A more strictly incremental detection should improve responsiveness (time-to-detection) too.

The clear omission in state-of-the-art systems is repair classification. We assume dialogue participants are sensitive to the function of a repair for several reasons. One may make direct use of semantic dependencies in substitutions such as “I saw [one, + {no,} two] men”, but may draw more pragmatic and turn taking inferences about utterance-initial deletes (restarts). Also, recognizing whether your dialogue partner sits either side of the statistically significant divide between “repeaters” and “deleters” (Shriberg, 1996) may help alignment. Classification’s obfuscation in the standard NLP disfluency task is perhaps due to its lack of clarity in definition. Verbatim repeats withstanding, as mentioned above there is often discrepancy between human annotations, suggesting gradient effects; finding a system that can reliably classify the extent of the repair and its function incrementally is a difficult challenge.

Given the problems with the various approaches, we are motivated to find a psychologically plausible incremental method for processing speech repair types by considering the time-linear order in which listeners receive the incoming acoustic signal and then react:

1. Detection of the *interruption point*, triggered via some combination of a partial word, an editing term forming an interregnum or characteristics of the repair onset.
2. Estimation of *reparandum start* position through some backward-looking process.
3. Possibly simultaneously with (2), estimation of the *repair end*, via detection of a further repair, a fluent continuation or the end of the utterance; interleaved with (1) and (2), the repair’s *classification*.

In the remainder of this paper we present a corpus survey in section 4 and a proposed approach for modelling repair in section 5 investigating these stages.

4 Self-Repair Distributions in Switchboard

Our initial repair distribution study uses the standard Switchboard training corpora (all conversation numbers sw2*,sw3* in the Penn Treebank III release), plus the non-Treebank Switchboard files, giving a total of 972 transcripts, $\sim$196,600 utterances, $\sim$1.28M words, from which we extract 40,485 self-repairs based on the annotations.

The base-rate likelihood of a given word beginning a repair onset is p= 0.0366, that is on average once every 27.3 words of speech.[2] We do not distinguish between repairs crossing utterance boundaries and those marked within an utterance unit, treating them both as first-position within one continual stream, however the difference between these two types would be interesting to consider in further study.

Repair taxonomy by alignment To investigate the distribution of the different types of repair, we follow Johnson and Charniak (2004) in their use of minimum string-edit distance alignment. Ignoring a handful of backwards-looking disfluencies which are annotated within editing term sequences, our aligner classifies 40,364 examples. It operates by mapping each reparandum word to a repair word, where each word must receive at least one alignment with the best possible score. In addition to their alignment categories we introduce COMPLETE_PARTIAL, which aligns prefix→complete word relations such as “j- + just”. We used the following scores to ensure that ‘weaker’ substitutional relations are replaced by stronger ones: REPEAT:6, COMPLETE_PARTIAL:5, SUB[same POS]:4, SUB[same POS first letter]:3, SUB[arbitrary]:2, DELETE:1 and INSERT:1. We decided that as COMPLETE_PARTIAL is a partial repeat it should be selected as a stronger alignment over a SUB[same POS].

The most frequent aligned structures extracted are shown in Table 1: we split the structures between the broad classes of verbatim repeats, substitutions and pure deletes (no repair phase annotated), in order to get the most prototypical deletes as judged by the Switchboard annotators.

1139 different alignment sequence types were found, with only 38.9% of types occurring at least twice, a figure higher than Heeman and Allen (1999)’s reported 29.4%, most likely due to a bigger corpus size. As can be seen, the majority of types are within substitutions, which have a long tail of compound types – the 10 example substitutions shown only constitute roughly half of all substitution occurrences. Deletes were the rarest,

---

[2] We exclude the first word of every utterance that is not a continuation, as you cannot begin a disfluency repair initiation across these boundaries. $\sim$100 repairs’ repair onset occur at the same word as an embedded repair, so simply dividing the number of word transitions by the number of repairs annotated would give a slight, but insignificant, boost in raw likelihood.

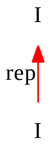
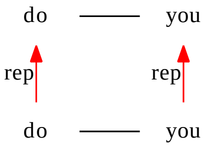
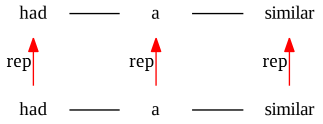
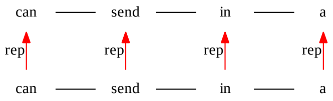
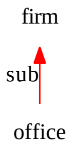
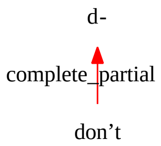
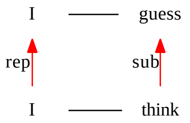
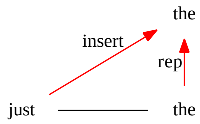
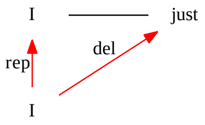
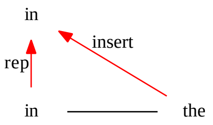
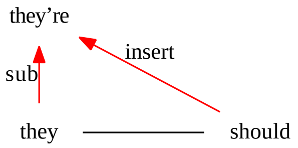
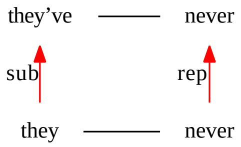
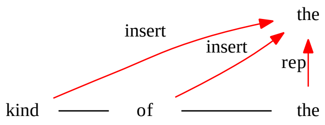
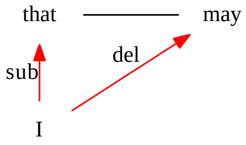
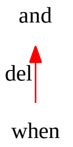
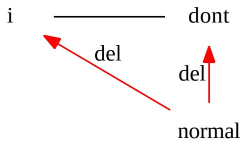
Repair class	Most Frequent repair types (% overall repairs)
Repeats (56.79%) +interregnum=11.96% of class; reparandum=1.23 (std=0.53, power $y = 1.7229x^{-4.425}$, $R^2 = 0.9565$);	 46.2%  8.2 %  had a similar rep rep rep had a similar 1.5 %  can send in a rep rep rep rep can send in a 0.3%
Substitutions (36.55%) +interregnum=18.65% of class; reparandum=1.78 (std=1.16, power $y = 1.0454x^{-2.593}$, $R^2 = 0.9227$);	 10.2%  3.3%  1.8%  1.4%  1.3%  1.0 %  0.9%  0.9 %  0.7%  0.7%
Deletes (6.66%) +interregnum=0.7% of class; reparandum=1.35 (std=0.88, power $y = 0.938x^{-2.995}$,$R^2 = 0.9956$);	 5.0%  0.8%

Table 1: Distribution of the most frequent repair disfluencies in Switchboard

conflicting with Shriberg (1996), but mainly due to our definition covering pure deletes only.

While building a rule-based repair grammar is not what we advocate in this paper, it is worth noting the observed alignment sequences can be compressed into 194 different operation sequence pairs such as [*SUB*(r_{m-i}-R_{n-j}) *REP* (r_m-R_n)], in this case representing a substitution alignment from i words back from current reparandum index m to a repair word j words back from current repair index n, followed by a repetition alignment between the current indices. In terms of coverage, due to the sparsity of most alignment sequences, the strength of Johnson and Charniak (2004)'s generative S-TAG grammar approach over a template based one (Heeman and Allen, 1999) becomes clear – for example the approach allows the most frequent repair type, repeats, to have high likelihood within a repair 'grammar', regardless of their length.

Reparandum lengths First-turn repairs tend to be very short, with a mean reparandum length of 1.44 (partial) words (pop. st.dev = 0.88). As with many linguistic phenomena, their length distribution can be characterized as an inverse power law: a function $y = 1.7197x^{-3.61}$, where x is the reparandum length in words and y is the average relative frequency of that length, has a goodness-of-fit R^2 = 0.9635 up to length 9. Reparanda of 1 or 2 words account for 90.8% of repairs and lengths 1-3 account for 96.5%. Repeats (1.23 words) and deletes (1.35 words) are significantly shorter than substitutions (1.78 words), which also exhibit a shallower power-law decay – see Table 1 for the figures.

With the vast majority of reparanda being 1-3 words long, a very local model of context could be used to capture them. As mentioned, previous approaches using sequence-based language models in combination with repair grammars and templates have had success, but there is scope for incorporating repair detection more directly into an n-gram model (though not necessarily through Hidden Event Language Models (HELMs) (Georgila, 2009), which require longer contexts and more training data). Furthermore, as Shriberg and Stolcke (1998) showed, the likelihood of retracing back one more word in retraces decays logarithmically with the number of words into a fluent word sequence, so the need to store all possible reparandum sites before having heard an interruption point seems unnecessarily complex: a locally triggered recovery mechanism does not have far to backtrack. Repeats and deletes are frequently short so their repair onset and reparanda will often fall within a bi- or trigram: for example, presuming perfect interregnum and edit term recognition, a trivial repeat-word feature $w_i = w_{i-1}$ captures 46.2% of all repairs. Use of such local alignments may yield high precision, but we need a more general way of detecting interruption points in a local n-gram context which can also capture longer repairs, as will be discussed below.

Embedded repairs 11.9% of all repairs are embedded inside a longer structure – this divides between 9.9% chaining repairs, embedded within the reparandum phase as in (6), and 2.0% nested within the repair phase of a longer repair.[3] While these appear to need more complex resolution mechanisms, which is presumably why they are ignored in the training phase and evaluation of automatic disfluency systems, they need not be processed as hierarchically embedded structures by listeners on-line. They are frequently short, with mean reparandum 1.28 words long (std=0.67), and so can be resolved very locally, again in a short n-gram context, and may provide an immediate feature for following repair onsets. Intuitively an interruption point indicates speaker trouble, so the likelihood of a consequent interruption point in the following word transitions increases.

(6) " [[This, + it,] + they] are really. "
Embedded chaining substitution- (sw3389)

Partial words as interruption point indicators The most reliable lexical indicator of a repair onset is a preceding partial word. According to the transcripts, the likelihood of a repair onset following a partial word that is not utterance-final is 0.925, boosting the likelihood significantly more than the presence of an interregnum, as will be discussed below. Furthermore, the remaining 0.075 of probability mass for continuations, upon inspection, look like mis-transcriptions. Reparandum-final partial words are present in 10.4% of repairs. Furthermore, the completion of a single partial word is one of the most frequent repair structures (3.3% of all repairs). The probability of the partial word

---

[3] While Shriberg (1994)'s thesis and Meteer and Taylor (1995)'s annotation attempted to formalise these, they remain a problem for consistency of annotation- it is not always clear whether they should be annotated as nested or chaining.

being a deleted reparandum also rises from the overall average rate 0.066 to 0.171.

This is clearly a very useful feature for detection and classification. Charniak and Johnson (2001) posit an optional phase between the reparandum and the interregnum called the 'free-final', consisting of a sequence of partial words of any length, which, when used as a training feature for an edited words classifier, can improve the detection of repairs. Subsequent work does not use partial words in an attempt to simulate a more realistic testing situation for dialogue systems. While we cannot make direct predictions here without the acoustic data, we investigate how a simple word completion predictor could be a fair approximation to an annotator's incremental processing in section 5.

Interregnum vocabulary Another incremental indication of repair, which has been established in previous empirical work (Clark and Fox Tree, 2002) and in formal models of dialogue (Ginzburg, 2012), is the presence of a conventional editing term for signalling speaker trouble. The editing signals that constitute most repair interregna have a characteristic vocabulary, a fact Heeman and Allen (1999)'s system exploited to detect them with almost perfect accuracy.

In Switchboard, only 13.9% of revision repairs have an interregnum, so it is not a strong repair indicator, which is surprising given its important role in formal and empirical models. However, if one is identified correctly, its presence signals information about the type of upcoming repair: the likelihood of a substitution rises to 0.499, and the likelihood of a delete reduces to $<$0.01, which could be due to deletion's more destructive semantic 'cancelling' function on the reparandum. There are more substitutions with interregna than repeats in raw frequency and significantly more relative to their class size (2752/14755 (18.65%), versus 2741/22921 (11.96%) $\chi^2_{(1)}$=322.9,p$<$0.0001).

Interregna share a virtually identical vocabulary to editing signals in the more common *abridged* (Heeman and Allen, 1999) or *forward-looking* (Ginzburg, 2012) repairs which comprise an editing signal followed by a fluent continuation to their preceding context, rather than a disfluent one. Focussing here on interregnum vocabulary distributions, we obtain the probabilities in the below table, showing the predictive power of the vocabulary item and its relative frequency within all repairs. The filled pause "uh" and discourse marker "you know" are the most indicative, increasing the probability of a repair from the base rate to 0.155 and 0.1 respectively. These two items are also the most frequently occurring within repairs (9.0% and 2.6% of repairs have them, respectively). The lack of predictive power even the most frequent interregna forms have to predict repair means interregnum presence does not provide a reliable feature for detection on its own; however as it has significant interaction with repair type, it is a useful feature for repair classification.

form	p(repair|form)	p(form|repair)
(fluent word)	0.037	0.861
"uh"	0.155	0.090
"you know"	0.100	0.026
"well"	0.080	0.006
"I mean"	0.074	0.005
"um"	0.061	0.003
"yeah"	0.038	0.002
"or"	0.017	0.002
"like"	0.014	0.003
"so"	0.005	0.001
"actually"	0.025	0.001

5 Language models for on-line repair processing

Having observed some distributional properties of the form of self-repairs that could contribute to on-line detection and classification tasks, we now introduce a simple information theoretic model which incorporates some of them, including local repair detection based on language model probability and partial word presence. This model can be used orthogonally to alignment approaches discussed above, and should provide scope for more efficient, realistic and robust implementations.

We model the task of listeners and annotators as representing the following constituents of a repair, ignoring interregna and other editing terms:

$$...w_o^N\,[w_{rm}^1...w_{rm}^N + w_{rp}^1...w_{rp}^N]w_c^1... \qquad (7)$$

Intuitively and in accordance with the processing order outlined in section 3, the first detection problem is recognizing the repair onset w_{rp}^1 (or w_c^1 for deletes). For this we intuit the most important factor is syntactic disfluency, that is, violation of syntactic expectation. Following a detection of this violation, the task is to find the start of the reparandum – which can be seen as maximising the fluency of a sequence including $w_o^N w_{rp}^1$ – while simultaneously beginning to compute the repair's parallelism to the reparandum onset w_{rm}^1. The final task is to find the repair end w_{rp}^N (or w_c^1 for deletes) and classify the repair through computing its parallelism to the reparandum up to its end w_{rm}^N. We discuss the tools we use to model violation of expectation and parallelism below.

Fluency measures for incremental repair onset detection We require a language model that can predict which word, or class of words, hearers are likely to hear next in on-going dialogue. Although we currently lack robust large-scale predictive incremental parsers – though see (Eshghi et al., 2013; Demberg et al., 2013) for on-going efforts – we can use an approximation to incremental lexical and syntactic fluency with n-gram language models and insights from recent work on modelling grammaticality judgements (Clark et al., 2013). We train a trigram model with Kneser-Ney smoothing (Kneser and Ney, 1995) as our principal default fluency measurement p^f:[4]

$$p^f(w_i \mid w_{i-2}, w_{i-1}) = p^{KN}(w_i \mid w_{i-2}, w_{i-1}) \quad (8)$$

We can define an additional measure of fluency based on the insights of the frequency of partial words at interruption points in section 4. We train a simple word completion model $p^{complete}(w|w_i)$ which operates on any annotated partial word prefix w_i to provide a distribution over possible completions, and thus the most likely completion (based on the prefix and unigram co-occurrence). For detection purposes, we make the realistic assumption that w_i can only be interpreted as an abandoned partial word after having encountered the following word w_{i+1}, which as the corpus study suggested is almost certain to be a repair onset w_{rp}^1. As opposed to leaving the partial word as unknown vocabulary we can instead define a probability distribution of the completion probability of each word in the vocabulary. So for a partial word w_i, the likelihood of w being its corresponding complete word at the time of interruption is:

$$p^{fluent}(w \mid w_{i-2}, w_{i-1}, w_i) = \frac{1}{Z} \times p^{KN}(w \mid w_{i-2}, w_{i-1}) \times p^{complete}(w \mid w_i) \quad (9)$$

where Z is a standard normalisation constant to ensure that $\sum_{w \in Vocab} p^{fluent}(w \mid w_{i-2}, w_{i-1}, w_i) = 1$. The probability $p^{\hat{fluent}}$ of most likely completion of w_i is then:

$$p^{\hat{fluent}} = \max_w p^{fluent}(w \mid w_{i-2}, w_{i-1}, w_i) \quad (10)$$

The intuition here is that when they encounter a partial word hearers attempt to find the most likely fluent word that both maximises its likelihood to be its completion and also of being a continuation of the two preceding words. If we encounter "yes I remem-", the probability of the completer's best guess will not be as low as if it was unpredictable, such as after an utterance initial "T-". When w_i is partial we use $p^{\hat{fluent}}$ in (10) for our fluency measure p^f, otherwise defaulting to our normal p^{KN} model.

Syntactic fluency measures Use of a standard n-gram model conflates syntactic with lexical predictability. To remove lexical effects and focus on syntactic effects only, we normalise for lexical probabilities by following Clark et al. (2013)'s use of Weighted Mean Logprob (WML). WML divides the logprob of the raw probabilities of all the trigrams in the utterance so far over the summed logprob of the component unigrams, normalising by the length of the utterance so far. We intend to use this incrementally and within local trigram windows rather than for full utterances. So at word w_i, we define our syntactic fluency measure as:

$$WML(w_{i-2} \ldots w_i) = \frac{1}{n} \frac{logp^f_{TRIGRAM}(\langle w_{i-2} \ldots w_i \rangle)}{logp^f_{UNIGRAM}(\langle w_{i-2} \ldots w_i \rangle)} \quad (11)$$

Repair classification by entropy measurement If a low WML measure or low p^f can indicate disfluency, a listener or annotator would then want to compute how similar two contexts were in order to infer the class of repair. To do this using trigram contexts we need a distribution of continuations after each word in repair utterances to be available, which we will refer to as $\theta^f(w \mid w_{i-1}, w_i)$. We can then take the entropy $H(\theta^f)$ to give us a measure of uncertainty in the distribution.

To measure syntactic and lexical parallelism between two words we measure the Kullback-Leibler (KL) divergence (relative entropy) between two different distributions of θ^f. This measure of parallelism will be particularly useful for

---

[4] Many thanks for use of the excellent code attached to Clark et al's paper, available for download at `http://www.dcs.kcl.ac.uk/staff/lappin/smog/`

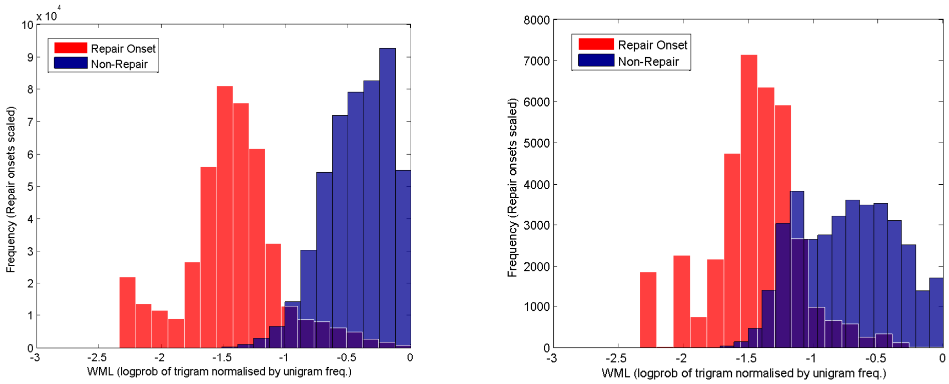


Figure 1: WML fluency measure for training data (left) and heldout data (right)

classification when comparing the θ^f of reparandum and repair boundary words, as will be explained below.

Hypotheses For the incremental processing of self-repair detection and classification, in terms of our fluency and parallelism measures, we hypothesise the following:

1. **Detection**: Repair onsets w_{rp}^1 with their context will have significantly lower mean p^f values than non-repair transition trigrams (lower lexical-syntactic probability), and exhibit considerably bigger drops in WML (lower syntactic probability) than other fluent trigrams in the utterance so far, caused by a partial word followed by a fluent one, or other syntactic disfluency.
2. **Reparandum start identification** Processing the utterance with the reparandum removed appropriately will significantly increase the WML of the utterance so far (similar intuition to the noisy channel approach), more so than other hypotheses for w_{rp}^1.
3. **Classification** For repeats, the KL divergence from the continuation distribution after the reparandum's first word, i.e. $\theta^f(w \mid w_o^N, w_{rm}^1)$, and that of the repair onset and its cleaned context before the reparandum, i.e. $\theta^f(w \mid w_o^N, w_{rp}^1)$, will trivially be 0 in repeats and repeat-initiated substitutions, will be greater for other substitutions and higher still for deletes.
4. **Partial word repair classification** We predict repairs with reparandum-final partial words w_{rm}^N with high entropy over possible completions θ^{fluent} (see equation (9)) will be interpreted as deletes rather than substitutions- in deletes the high uncertainty of predicted complete word is interpreted as 'cancelled'.
5. **Repair end detection/final classification**: In repeats, the continuation distribution at the reparandum-final word w_{rm}^N (i.e. $\theta^f(w \mid w_{rm}^{N-1}, w_{rm}^N)$) will be maximally close to that at the repair-final word w_{rp}^N (i.e. $\theta^f(w \mid w_{rp}^{N-1}, w_{rp}^N)$) with KL divergence 0. In substitutions, the same KL divergence will be on average higher than in repeats (though for compound type repairs ending in repeats this could still be 0), and the KL divergence for deletes should be even higher.[5] Substitutions as a class may vary significantly within this measure and in the KL divergence in hypothesis (3), however one KL divergence should be sufficiently lower than that of an average delete, and one should be higher than 0 due to them not being verbatim repeats.

Experiments At the time of writing we have investigated hypothesis (1) using the standard division for the Switchboard disfluency detection task for training and held-out data (Charniak and Johnson, 2001, *inter alia*),[6] and for now omitting partial words as per the normal task. After training on a cleaned model (reparandum and edit-terms excised) from the standard Switchboard training data (100K utterances, 650K words), which when

---

[5] We approximate divergence between $\theta^f(w \mid w_{rm}^{N-1}, w_{rm}^N)$ and $\theta^f(w \mid w_{rm}^N, w_{rp}^1)$ in deletes, due to the lack of a repair phase; the distribution of continuations after the repair onset (first non-reparandum word) is our best approximation of the repair end.

[6] We reserve the normal test data files for future work.

run over the same training corpus with disfluencies included the model assigns a mean WML of -0.432 (std.=0.262) to non-repair onset trigrams and -1.434 (std.=0.388) to repair onsets. Encouragingly on unseen data, the standard held-out data (PTB III sw4[5-9]*, 6.4K utterances, 49K words.) there is still a significant difference: fluent trigrams had a lower mean, -0.736 (sd=0.359) while repair onsets were similar to their training average at -1.457 (std.=0.359)– see Figure 1. We suspect the sparsity of clean data may have caused this shift, so we would expect to see the effect maintain a healthy gap in testing with a larger language models.

6 Discussion

We have described self-repair processing in terms of probabilistic expectation violation and distributional distance in a fluent language model. We argue this could be a more realistic model than alignment driven self-repair detection posited in state-of-the-art computational models, due to its efficiency and lack of over-prediction. The repair onset detection can be triggered with no latency through using a simple language model. We hope to show conclusively in future work that the many different types of repair distinguished by automatic alignment in our corpus study can be captured by our simple information-theoretic model of incremental fluency estimation and local repair.

Acknowledgements Thanks to the SemDial reviewers for their helpful comments and to Shalom Lappin and Gianluca Giorgolo for their code.

References

S. Brennan and M. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

E. Charniak and M. Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the second meeting of the NAACL on Language technologies*, ACL.

H. H. Clark and J. E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111.

A. Clark, G. Giorgolo, and S. Lappin. 2013. Statistical representation of grammaticality judgements: the limits of n-gram models. In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. Bulgaria. ACL

V. Demberg, F. Keller, and A. Koller. 2013. Parsing with psycholinguistically motivated tree-adjoining grammar. *to appear in Computational Linguistics, Vol. 39, No. 4*.

A. Eshghi, J. Hough, and M. Purver. 2013. Incremental grammar induction from child-directed dialogue utterances. In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. Bulgaria. ACL

K. Georgila. 2009. Using integer linear programming for detecting speech disfluencies. In *Proceedings of Human Language Technologies: NAACL 2009*, pages 109–112. ACL

J. Ginzburg, R. Fernández, and D. Schlangen. 2013. Dysfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*.

J. Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

P. Heeman and J. Allen. 1999. Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

M. Johnson and E. Charniak. 2004. A tag-based noisy channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the ACL*, ACL '04.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *ICASSP-95*, volume 1, pages 181–184. IEEE.

W. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

M. Meteer and A. Taylor. 1995. Disfluency annotation stylebook for the switchboard corpus. ms. Technical report, Department of Computer and Information Science, University of Pennsylvania.

X. Qian and Y. Liu. 2013. Disfluency detection using multi-step stacked learning. In *Proceedings of NAACL-HLT*, pages 820–825.

E. Shriberg and A. Stolcke. 1998. How far do speakers back up in repairs? A quantitative model. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2183–2186.

E. Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley.

E. Shriberg. 1996. Disfluencies in switchboard. In *In Proceedings of the ICSLP 96*, volume 96, pages 3–6. Citeseer.

S. Zwarts, M. Johnson, and R. Dale. 2010. Detecting speech repairs incrementally using a noisy channel approach. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10. ACL.

Be Open To Computer Based Coaching

Daniel Mäurer and **Karsten Weihe**
Departement of Computer Science
Darmstadt University of Technology
Hochschulstr. 10 64283 Darmstadt
`{maeurer,weihe}@algo.informatik.tu-darmstadt.de`

Anna Bruns
Institute of Psychology
Darmstadt University of Technology
Alexanderstr. 10 64283 Darmstadt
`bruns@psych.tu-darmstadt.de`

Abstract

Conversational agents are used for various purposes, such as marketing, e-learning and tutoring. However, they have not been used for personal coaching so far. Personal coaching can be used as a strategy to support professionals in transferring their newly acquired skills to everyday work after receiving a training. The aim of our research was to examine the usefulness of computer based coaching as a training transfer strategy. We also examined whether the user's openness is a key factor for the effectiveness of a virtual coach. We present a computer based coaching system specifically designed for training transfer. In a longitudinal experiment we tested our system against an online journal with regard to its effectiveness. We found some evidence that participants with high openness benefit more from computer based coaching than participants with lower openness, while openness has no influence on the effectiveness of the online journal. Our results suggest that computer based coaching can be effectively used as a training transfer strategy, but may not be equally effective for everyone.

1 Introduction

Conversational agents are used for marketing, e-learning and tutoring purposes, just to mention a few. Whereas users remain skeptical towards "talking" to a computer, there are still many unexplored application areas for conversational agents. Personal coaching, for example, is a promising field as the coaching maxim rather involves the guidance of individuals by stimulating questions than by specific instructions or correct answers. So far, conversational agents have not been extensively used for personal coaching purposes.

We present a computer based coaching system that is particularly designed to enhance training transfer. Training transfer, the application of newly acquired skills in everyday work, requires a maximum of support and reflection. This support and reflection can be provided by a personal coach (Olivero et al., 1997). However, personal coaching is very expensive and therefore only available to a limited circle of individuals in organizations. Computer based coaching, on the other hand, could support a much larger audience to successfully transfer their newly acquired skills from training to everyday work.

Our Contribution We have developed a conversational agent (a.k.a. dialog system) to support our computer based communication training. The dialog system was accessible for the participants during weekdays to support them with the application of acquired communication skills after a weekend training phase. This kind of computer based coaching is intended to support trainees to reflect on their process of goal accomplishment by targeted questions. Thereby, it does not understand the full content of the trainees' reactions to this questions in detail. Instead, the system classifies trainees' reactions as dialog acts (Stolcke et al., 2000) to keep track of the relevant information to successfully direct and control the dialog process.

Of course, our goal is not to replace real human coaches in general; just like Weizenbaum (1966) was sure that ELIZA is not a way to replace psychotherapists. Our vision is to introduce computer based coaching in situations where a personal human coach is simply not affordable or available.

Ever since, user acceptance has been a serious problem for conversational agents. In a professional setting such as training transfer, user acceptance is a key success factor. As the personality trait openness of the coachees is known to affect user acceptance (Devaraj et al., 2008), we expect that it can also have a huge impact on the transfer success as well. In line with this, we also expect that our results will stimulate a new direction of research on computer based coaching and the influence of personality traits.

Testing Effectiveness In order to test the effectiveness of our computer based coaching system, we present a longitudinal field experimental study that compares two training transfer support strategies after an online communication training: computer based coaching and online journals. This experiment is expected to reveal first results on the effectiveness of computer based coaching as training transfer strategy. Furthermore, we want to explore how openness and user acceptance interact with the effectiveness of computer based coaching. Our results can provide a basis for future developments of conversational agents in the field of coaching.

Related Work Many other work has been done in the field of conversational agents, but only few focused on coaching. The results closest to ours are found in the healthcare domain, where (Bickmore et al., 2005) present an agent with the role of a physical exercise advisor. Although their dialog manager is working in a similar way to the one we present, user contributions to the dialog are made primarily by selecting items from multiple-choice menus, whereas our system allows the user to answer in natural language at every time. While multiple-choice menus are a sufficient way to enforce the users motivation for physical exercising, we expect communication in natural language to be necessary for cognitive tasks such as training transfer. (Bickmore, 2003) focused on studying the effects of social conversation and the relation between the agent and the user in a artificial situation. He also found that the user personality and their trust in the agent were intercorrelated. In contrast, our study focuses on the outcome on the domain goal in a real world setting, namely the increase in their communication skills. SimCoach, a dialog based healthcare assistant, (Rizzo et al., 2011) focused on promoting access to domain specific information. (Conati et al., 2000) introduced a chat based tutor in an educational setting. This tutor aims to foster learning from examples and to provide feedback on self-testing examples.

This paper is structured as follows. Section 2 will introduce coaching and the application of coaching as a training transfer strategy. In Section 3, we will introduce our computer based coaching agent. Section 4 will describe our experimental study on the effectiveness of our system. We will finish with the results of the study in Section 5 and an outlook on our future work.

2 Coaching

2.1 Solution Based Brief Coaching

Like many other disciplines, coaching has struggled with developing a common definition. For the scenario of a short training transfer dialog, we picked a goal-focused approach called *brief coaching* (Berg and Szabo, 2005). The primary method of solution based *brief coaching* is to support the coachee in defining goals and a suitable goal accomplishment strategy. A coaching session in *brief coaching* encompasses three stage phases:

- **Desired Future:** Defining a specific goal.
- **Changes in State:** Discuss recent past, look for indicators of changes in direction of desired state.
- **Experimental Phase:** Put into practice what has been discussed so far, agree on minor changes in everyday activity.

A general maxim of many coaching approaches is to regard the client as the expert for the relevant problem rather than seeing the coach as the expert for the client's problem. Therefore, coaching does not intend to give advice or push the client into any certain direction, but rather to ask targeted questions that help the 'expert' to get a new perspective and to develop his own solution. An ideal coach would do this by stating questions *only*.

2.2 Coaching as a Training Transfer Strategy

Training is successful, if training transfer was successful (Barnett and Ceci, 2002). Training transfer is defined as "...the degree to which trainees effectively apply the knowledge, skills, and attitudes gained in a training context to the job. For transfer to have occurred, learned behavior must be generalized to the job context and maintained over a period of time on the job." (Baldwin and Ford, 1988, p. 63).

However, at work, costs of failure and the pressure to meet deadlines typically hinder employees from exploring new and alternative methods. Instead of further improving their skills, they tend to rely on existing and well-practiced methods (Ericsson et al., 1993; Haccoun, 1997). Transfer interventions are effectively used to increase the motivation of learners to use their newly acquired skills in their daily routine. In particular, literature strongly supports the use of the goal setting strategy (Burke and Hutchins, 2007). Comparative studies reveal that goal setting is superior to other post-training interventions in terms of increasing trainees' transfer performance (Wexley and Baldwin, 1986). As solution-based brief coaching (Berg and Szabo, 2005) is a particular form of goal setting, we suggest coaching to be a suitable alternative to conventional post training interventions (e.g. goal setting via online journals). Decisive superiority of coaching in comparison to other transfer strategies may lie in its ability to enhance participants' metacognitions (Grant, 2003). Metacognitions capture the planning how to best achieve a specific goal, monitoring the progress and the evaluation of the used strategies (Schraw and Moshman, 1995). First research attempts were able to show that managers who received personal coaching after a training intervention further increased their productivity during the coaching phase (Olivero et al., 1997). However, personal coaching would be far too expensive in order to provide it to a larger number of employees.

Our computer based coaching , on the contrary, could be an effective and economic alternative.

Openness We expect personality to affect the success of computer based coaching. "Openness to new experience" is one of the "Big Five" personality factors. It encompasses intellectual curiosity, preference for variety and the willingness to explore new ways (Costa and McCrae, 1992). Especially for complex and changing task conditions, openness has shown to significantly impact the effectiveness of training interventions (Herold et al., 2002).

Furthermore, openness has a significant positive influence on coaching success (Stewart et al., 2008) and certain components of user acceptance (Devaraj et al., 2008). In contrast to conventional transfer strategies (e.g. online journals), we expect user acceptance to be a key factor for the success of computer based coaching. Therefore, we argue that individuals with higher openness will benefit more from computer based coaching than individuals with low openness.

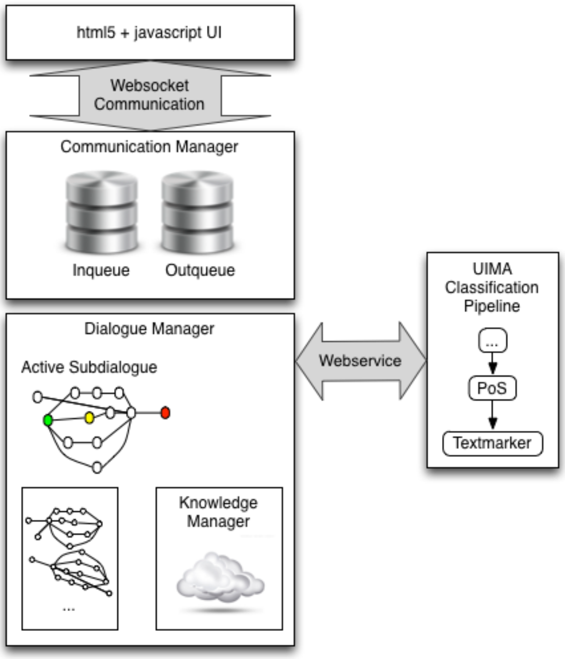


Figure 1: System Architecture

3 A chat based coaching system

In this section we will describe some of the details of our computer based coaching system. It is a mixed initiative system, which means that both the user and our system can take initiative turns to start a subdialog. A turn is a single utterance either by the user or the system. Although our system is technically capable of handling user initiatives as well, the coach is the one who usually takes the initiative in an ideal solution-based brief coaching process. Our conversational agent works with a high level dialog structure and local subdialogs. This high level structure follows the three stages model of solution-based brief coaching.

Additionally, we provide a set of transition rules that determine which subdialog is selected depending on the user's progress, for example:
if $worked\_on\_goals\_today?$
then $subdialog$: $''percentage?''$
else $subdialog$: $''what\_went\_wrong?''$

Our system can be divided into several components (Figure 1), namely *dialog manager*, *subdialog system*, *text classification* and *communication manager*. We will shortly describe them in the following:

Dialog Manager The main function of the dialog manager is to control the flow of the dialog. This means that it keeps track of the context knowledge, state of the conversation and all knowledge gathered from further sessions. Every single coaching question defined in the high level conversation structure is representing one subdialog. The system responses are constructed from a large set of pre-formulated sentences or sentence snippets. This also provides several alternative formulations for each possible system response.

After finishing a subdialog, a specific content analyzation component is extracting relevant changes in knowledge and forwards them to the knowledge manager. The analyzation component also decides whether the dialog manager should publish some feedback responses to the user.

Subdialog System The purpose of a subdialog is to extract a unique piece of information from the user. Each subdialog is based on a finite state machine. Transitions between the different states are triggered by user or system dialog acts. The users utterances are classified by a rule based classifier as described in the next paragraph. Each subdialog models a coaching question with all turns required to answer the question sufficiently. In an ideal case, this subdialog can be closed in two turns, for example a *y/n-question* can be simply responded with an *agreement* dialog act. A more complicated example would be the following which was taken from a real dialog (See appendix).
The coach is marked 'C', the human user with 'H'. In parentheses: The dialog act of the utterance.

...
***C:** Ok. How would you rate yesterdays progress on your weekly goal? Did you fully reach your goal or only to some extent?* (question)
***H:** Yesterday, one my supervisor talked to me about a mistake I made.* (statement)
***C:** Go on.* (pull)
***H:** I remained calm and analysed the situation. Based on my analysis I recognized that my mistake can be corrected.* (statement)
***C:** And how would you rate your goal accomplishment in percent? (0 - 100)?* (question)
***H:** I have reached my goal by about 80%.* (statement)
***C:** Congratulations, well done.* (acknowledge)
...

In this subdialog the system asks for the amount of progress. With his first response, the user does not answer this question properly and responds with a loosely related statement instead. The system cannot handle the information correctly, but this is not obvious to the user. Instead, after another turn the system simply tries to restate the initial question and asks for the percentage of progress. When reaching the final state and analyzing the result, the system ends the subdialog with a feedback response.

Our system also provides fallback mechanisms in case of mismatches or unrecognized turns made by the user. An example for a fallback is a handler for counter questions, which suspends the current subdialog and resumes it afterwards. Another handler is implemented for problem recognition, for example too many turns in one subdialog or too many unexpected user responses.

Text Classification: We decided to develop a rule based dialog act classifier. Other approaches, mostly for the English language, are machine learning or statistical approaches as presented in various publications (Stolcke et al., 2000; Marineau et al., 2000; Reithinger and Klesen, 1997). Machine learning was not an option because of non available suitable corpus for chat conversation in German language.

We implemented a UIMA[1] based classification pipeline using tokenization, lemmatization and a part-of-speech tagger for German language. Our set of classification rules at the end of the pipeline was implemented with TextMarker (Kluegl et al., 2009), now known as *UIMA ruta*.

Mixed Initiative Multi-turn Management: In order to improve the acceptance of the system we developed a communication manager protocol capable of multi-turn interactions. The protocol was specifically designed to simulate human chat behavior, for example that the user has the possibility to state more than one submission to the system:
...
***C:** How are you today?*
***H:** mhm...*
(5 sec pause)
***H:** not that good.*
***C:** Why is that?*
...
Similar to the user, our system is able to state more than one single utterance within a short amount

---

[1] UIMA stands for Unstructured Information Management Architecture and is a component architecture and software framework implementation for the analysis of unstructured content like text. See http://uima.apache.org for details.

of time. Therefore, our communication manager component works with two queue data structures. One queue is for the user utterances published since the system's last dialog act, called *incoming queue*. The other queue stores system utterances which the dialog manager has planned to publish to the user, but has not published yet. This one is called *outgoing queue*.

Once an utterance from the user is received, the dialog manager compiles a response and pushes it into the outgoing queue. These responses are prepared to be published under the given context. Depending on the question stated by the system and the new response to be posted, the system will determine a waiting time depending on the responses length and question initially stated and wait for that amount of time. While waiting, the user will be informed via the *typing* signal that the system is planning to say something. However, the user is still able to make another turn or add something to his/her last statement. If the user had something more to say, the system clears its outgoing queue with all the utterances the system planned to publish. The Dialog Manager recompiles its response based on the new state of information. Once the system is sure that the current response is valid under the context of all given answers by the user, it will publish its response and clears the ingoing queue with all the user's messages it had to respond to. With this protocol we improve naturalness and assure consistency in all system responses.

User Interface: Our chatcoach is presented via a web-based html 5 frontend. It was a very important requirement that our system is easily accessible. Our web frontend even enables support for mobile devices. The communication between the browser based client frontend and our server-based agent is handled by a publish/subscribe messaging framework. This gives the system the possibility to also display *typing* status messages which are very important for handling the message publication timing in a multiple user turn scenario. Every typing action of the user or of the system is immediately present to the other party. This enables a more lifelike user experience during the conversation.

Example Dialog The following dialog shows an example session of a workday translated from German.

C: *Hello. The last time we talked to each other you told me you couldn't work much on your goal. Did you come any closer in the meantime?*
H: *Yes, always whenever mistake occurs, I try to stay calm and analyse the situation.*
C: *Ok. How would you rate yesterdays progress on your weekly goal? Did you fully reach your goal or only to some extent?*
H: *Yesterday, one my supervisor talked to me about a mistake I made.*
C: *Go on.*
H: *I remained calm and analysed the situation. Based on my analysis I recognized that my mistake can be corrected.*
C: *And how would you rate your goal accomplishment in percent? (0 - 100)?*
H: *I have reached my goal by about 80%.*
C: *Congratulations, well done.*
C: *Why do you think you haven't accomplished more than that?*
H: *At first, it is always a shock and pretty difficult to stay calm. But when I take the time and analyse the situation, its getting better.*
C: *Is there anything that went particularly good, when you think about working on your goal yesterday?*
H: *Yes, that i remained calm although I was stressed.*
C: *What exactly was positive about that situation?*
H: *To stay calm and not get hectic*
C: *Ok good. Now that we have reflected what went well and not so well on your last workday is there something that you would like to change tomorrow?*
H: *No, I think its okay.*
H: *I'm starting to get a routine in such situations.*
C: *And what could help you reach your weekly goal completely?*
H: *I guess thats not really possible in the near future.*
C: *Are you satisfied with your actual goal accomplishment?*
H: *I think I am.*
C: *Sounds good.*
C: *And how much would you say does the work on your goal affect your communication skills? (strong to not at all)*
H: *I would say strong*
C: *Alright. I think that's it for today. Goodbye!*
H: *Bye*

	Total M (SD)	Journal M (SD)	cbc M (SD)	1	2	3	4	5	6
Transfer Strategy	-0.08 (1.02)	1.00 (0.00)	-1.00 (0.00)	-					
Communication Skills (T1)	19.86 (4.83)	19.94 (5.07)	19.84 (4.81)	-.01	-				
Communication Skills (T2)	20.68 (6.09)	20.28 (5.42)	21.02 (6.79)	.06	.40*	-			
Communication Skills (T3)	22.57 (4.32)	22.43 (3.34)	22.68 (5.14)	.03	.44*	.49*	-		
Openness	2.54 (1.07)	2.42 (1.38)	2.64 (0.74)	.11	-.07	.10	.03	–	
Motivation	2.20 (0.51)	2.26 (0.60)	2.15 (0.43)	-.10	.33	.31	.22	.18	-
Login Frequency	4.23 (4.62)	2.50 (1.73)	5.71 (5.77)	$.35^T$	-.17	-.08	-.28	.19	-.06

Table 1: Means, Standard Deviations and Pearson Correlations. Note: *Computer Based Coaching vs. Online Journal* was contrast coded: Computer Based Coaching (cbc) = 1; Online Journal (journal) = -1. $^T p < .10, ^* p < .05$.

4 Experimental Study

Our study was conducted as a longitudinal field experiment with three measurement points. Participants were 26 alumni of TU Darmstadt (50% female) who voluntarily signed up for an online communication training program. On average, participants were 36 years old (SD = 10.89) and completed one to five communication trainings prior to this training.

Our study was conducted in two phases: A pretest on participants' Communication Skills (t1) was followed by a training phase that ended with a posttest on Communication Skills (t2). Immediately after the posttest, participants were randomly assigned to two different experimental conditions: Twelve participants were instructed to record their progress in an online journal on a daily basis over the course of one week. The remaining fourteen participants were instructed to use the computer based coaching also on a daily basis over the course of one week. The effects of the different transfer strategies were assessed in a follow-up test on participants' Communication Skills (t3) after the end of the transfer week.

Instructions after the posttest (t2) were identical in both experimental conditions (Computer Based Coaching and Online Journal) in the following aspects: On the first day (after the posttest at t2) participants were asked to define a specific goal they want to accomplish in the transfer phase. This goal should refer the improvement of communication skills. Positive examples for goals were provided. Participants were also asked to rate the feasibility, and to name potential promoters and inhibitors to accomplish their goal.

On the following four days participants were instructed to rate goal accomplishment and the (positive or negative) effect of this goal accomplishment on their communication skills. Furthermore we asked participants to name the specific promoters and inhibitors of goal accomplishment they faced during that day.

Instructions after the posttest (t2) were different for both experimental conditions (Computer Based Coaching and Online Journal) in the following aspect only: Whereas the Online Journal presented the instructions in a static form, the Computer Based Coaching presented the questions adaptively in the form of a dialog as described in Section 3.

4.1 Measures

Communication skills were assessed in a test at all three measurement points: in a pretest prior to the training phase (t1), in a posttest after the training phase and prior to the experimental manipulation of the Transfer Strategies (t2) and in a follow-up after the experimental manipulation of the Transfer Strategies (t3). The test consisted of three critical situations that were presented to the participants (i.e. 9 critical situations in total). Within 15 minutes, participants had to generate as many useful and original responses to the given situations as possible. Two independent experts rated the quality of the different responses on two dimensions (usefulness and originality) on an anchored 7-point Likert scale. Multiple responses of an individual participant were averaged per situation and dimension. A single Communication Skills Index was formed by multiplying scores on these two dimensions (Zhou and Oldham, 2001). A global Communication Skills Index per measurement point was aggregated across the three test situations. This elaborate procedure resulted in a good agreement between the ratings of the two independent experts (ICC .70 to .84).

Openness was measured by two items derived from (Rammstedt and John, 2007) Big Five Inventory-10. Participants rated themselves on both items ('I see myself as someone who has as few artistic interests.' and 'I see myself as someone who has an active imagination.') on a five-point Likert scale (1= disagree strongly to 5 = agree strongly). Both items were later aggregated to a global Openness score.

4.2 Control Variables

In our analysis, we wanted to see the "pure" effect of our Transfer Strategy(Computer Based Coaching and Online Journal) without the distortion of other influential factors. Therefore we controlled for several variables in our analysis that we expected to also have an influence on Communication skills at t3 apart from our Transfer Strategy:

First, we expected our participants to differ in their Communications Skills prior to the training (at t1) and prior to the experimental manipulation of the Transfer Strategy (t2). In order to eliminate in our analysis both the influence of prior Communications Skills and the effects of the training itself, we included Communication Skills at t1 and t2 as control variables into our analysis.

Second, we provided participants with access to the training chapters also after they had completed the posttest at t2. As further repetition of the training chapters may also cause a further improvement of participants' Communication Skills, we recorded the Login Frequency after the posttest and controlled for its influence in our analysis.

Third, we expected the participants' individual motivation to have an influence on the effectiveness of the training and possibly interfere with the effects of the different Transfer Strategies. Therefore, we assessed participants' initial motivation to sign up for our communication training on 15 items (e.g. "My main driver to participate in the training is because I want to improve my social skills") that covered five dimensions of motivators from technical aspects to career advancement. Participants rated their motivation on a five-point Likert scale (1= disagree strongly to 5 = agree strongly). All items were later aggregated to a global motivation score.

5 Results

5.1 Descriptives

Our analysis encompassed one dependent variable (Communication Skills at t3), two independent variables (Openness and Transfer Strategy) and four Control Variables (Communication Skills at t1, Communication Skills at t2, Login Frequency and Motivation). Means, standard deviations, and intercorrelations among all variables are presented in Table 1. Communication Skills are significantly correlated ($p < .05$) across the three measurement points. Furthermore, the Transfer Strategy and Login Frequency were significantly correlated ($p < .10$). As Computer Based Coaching was contrast coded with +1 and the Online Journal with -1 this positive correlation indicates that participants in the *Computer Based Coaching* condition had more logins after the posttest at t2 than participants in the Online Journal condition. All other variables did not differ significantly between the two experimental conditions.

5.2 Hypothesis Testing

We assumed that participants who are more open to new experience will benefit more from computer based coaching than participants who are less open to new experience.

We tested our assumption using hierarchical regression analysis. In the first step, we entered the Control variables (Communication Skills at t1, Communications Skills at t2, Login Frequency and Motivation). In the second step, we entered the moderator variable (Openness) and Transfer Strategy (*Computer Based Coaching vs. Online Journal*). The interaction term between Transfer Strategy and Openness was entered in the third step (Aiken and West, 1991).

To reduce multicollinearity, all variables were centred at their respective means.

Table 2 reports the test of our assumption: The Control variables entered in step 1 of the hierar-

	Step 1	Step 2	Step 3
Change in variance accounted for Communication Skills at t3 (ΔR^2)	.36*	.01	.14*
Communication Skills (t1)	0.26	0.26	0.12
Communication Skills (t2)	0.37^T	0.36^T	0.50*
Login Frequency	-0.21	-0.25	-0.33^T
Motivation	0.01	0.01	0.28
Transfer Strat.	—	0.09	0.08
Openness	—	0.05	0.26
Transfer Strat. x Openness	—	—	0.55*

Table 2: Test of Moderation Transfer Strategy x Openness on Communication Skills. Note: Values are standardized regression coefficients. All terms were centred prior to analysis.
^{T} : $p < .10$ *:$p < .05$.

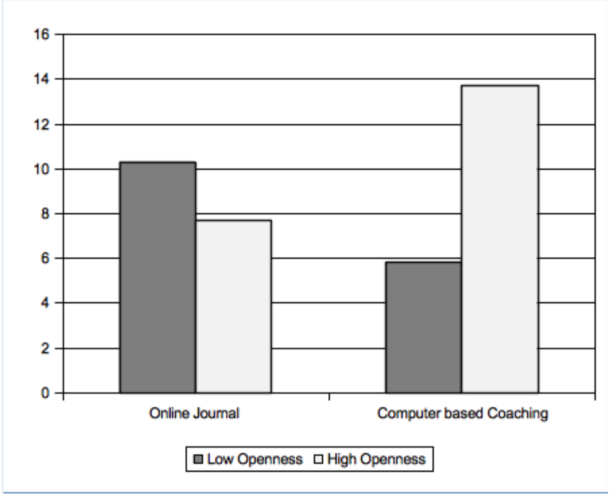


Figure 2: Gain in Communication Skills at t3 depending on Openness and Transfer Strategy after controlling for Communication Skills at t1, Communication Skills at t2, Login Frequency and Motivation.

chical regression analysis already accounted for 36% of the variance in the Communications Skills at t3. In particular, Communications Skills at t2 significantly predict Communications Skills at t3 ($\beta = .37, p < .10$). When we entered the Transfer Strategy (Online Journal vs. computer based coaching) and Openness in step 2 no additional variance in the Communications Skills at t3 was explained. However, when we entered the interaction between Transfer Strategy and Openness in Step 3 additional 14% of the variance in the Communications Skills at t3 could be explained. In sum, a total of 50% of the variance in the Communications Skills at t3 can be explained by using this set of variables. Among the variables three significant predictors were identified: The interaction between Transfer Strategy and Openness significantly predicted Communication Skills at t3 together with the Control variables Communication Skills at t2 ($\beta = .50, p < .05$) and Login Frequency ($\beta = -.33, p < .10$).

The plot of the relationship between Transfer Strategy and Openness is presented in Figure 2 and supported our Hypothesis: Participants who are more open to new experience benefit more from the Computer Based Coaching-condition than participants who are less open to new experience. The simple slope analysis revealed this difference to be significant ($p < .05$). In the Online Journal-condition, the effect seemed to be reversed. However, the simple slope analysis revealed this difference not to be significant ($ns.$).

5.3 Quality Evaluation:

Classification error rates did not vary significantly between the high and the low openness group. Fatal classification errors, such as mistaking a *disagreement* for an *agreement*, were not observed during our study. One of the shortcomings of our system was its deficient handling of counter questions. We counted four dialogs where the user aborted the conversation. In three of those conversations, the situations that caused the dialog to fail were initiated by user questions or false-positive questions.

6 Conclusion and future work

Our results suggest that computer based coaching effectively helps the participants to further increase their communication skills after a training intervention. According to (Shawar and Atwell, 2007), the best method to evaluate a conversational agent is to measure whether it achieves the service or task it was intended to. In this respect, our system performed quite well.

However, the participants' success largely differ with regard to their openness: Participants with high levels of openness benefit more from computer based coaching than participants with low levels of openness. In contrast, openness for experience does not seem to influence the effectiveness of online journals. This implies that computer based coaching is probably not suitable for everyone. Therefore, future work on similar research questions should take into account the influence of personality and background. It may be advisable to consider the users personality and background in order to avoid biased results in similar studies.

Of course, the dialog system will be further improved and is planned to be used for other application scenarios, for example decision coaching. Our future work includes building a German chat corpus with dialog act tags. We are planning to use it for further evaluation and improvements of the dialog act classifier.

References

[Aiken and West1991] Leona S Aiken and Stephen G West. 1991. *Multiple regression: Testing and interpreting interactions*. Sage.

[Baldwin and Ford1988] Timothy T Baldwin and J Kevin Ford. 1988. Transfer of training: A review and directions for future research. *Personnel psychology*, 41(1):63–105.

[Barnett and Ceci2002] Susan M Barnett and Stephen J Ceci. 2002. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, 128(4):612.

[Berg and Szabo2005] Insoo Kim Berg and Peter Szabo. 2005. *Brief coaching for lasting solutions*. WW Norton New York.

[Bickmore et al.2005] Timothy Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 59(1):21 – 30.

[Bickmore2003] Timothy Wallace Bickmore. 2003. *Relational agents: Effecting change through human-computer relationships*. Ph.D. thesis, Massachusetts Institute of Technology.

[Burke and Hutchins2007] Lisa A Burke and Holly M Hutchins. 2007. Training transfer: An integrative literature review. *Human resource development review*, 6(3):263–296.

[Conati et al.2000] Cristina Conati, Kurt Vanlehn, et al. 2000. Toward computer-based support of metacognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education (IJAIED)*, 11:389–415.

[Costa and McCrae1992] Paul T Costa and Robert R McCrae. 1992. Four ways five factors are basic. *Personality and individual differences*, 13(6):653–665.

[Devaraj et al.2008] Sarv Devaraj, Robert F Easley, and J Michael Crant. 2008. Research notehow does personality matter? relating the five-factor model to technology acceptance and use. *Information Systems Research*, 19(1):93–105.

[Ericsson et al.1993] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363.

[Grant2003] Anthony M Grant. 2003. The impact of life coaching on goal attainment, metacognition and mental health. *Social Behavior and Personality: an international journal*, 31(3):253–263.

[Haccoun1997] Robert R Haccoun. 1997. Transfer and retention: Let's do both and avoid dilemmas. *Applied Psychology*, 46(4):340–344.

[Herold et al.2002] David M Herold, Walter Davis, Donald B Fedor, and Charles K Parsons. 2002. Dispositional influences on transfer of learning in multistage training programs. *Personnel Psychology*, 55(4):851–869.

[Kluegl et al.2009] Peter Kluegl, Martin Atzmueller, and Frank Puppe. 2009. Textmarker: A tool for rule-based information extraction. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Proceedings of the Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop*, pages 233–240. Gunter Narr Verlag.

[Marineau et al.2000] J. Marineau, P. Wiemer-Hastings, D. Harter, B. Olde, P. Chipman, A. Karnavat, V. Pomeroy, A. Graesser, and the TRG. 2000. Classification of speech acts in tutorial dialog. In *Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference*, pages 65–71.

[Olivero et al.1997] Gerald Olivero, K Denise Bane, and Richard E Kopelman. 1997. Executive coaching as a transfer of training tool: Effects on productivity in a public agency. *Public personnel management*, 26:461–470.

[Rammstedt and John2007] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.

[Reithinger and Klesen1997] Norbert Reithinger and Martin Klesen. 1997. Dialogue act classification using language models. In *EuroSpeech*. Citeseer.

[Rizzo et al.2011] Albert Rizzo, K Sagae, E Forbell, J Kim, B Lange, JG Buckwalter, Josh Williams, TD Parsons, P Kenny, David Traum, et al. 2011. Simcoach: an intelligent virtual human system for providing healthcare information and support. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, volume 2011. NTSA.

[Schraw and Moshman1995] Gregory Schraw and David Moshman. 1995. Metacognitive theories. *Educational psychology review*, 7(4):351–371.

[Shawar and Atwell2007] Bayan Abu Shawar and Eric Atwell. 2007. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pages 89–96. Association for Computational Linguistics.

[Stewart et al.2008] Lorna J Stewart, Stephen Palmer, Helen Wilkin, and Marie Kerrin. 2008. The influence of character: Does personality impact coaching success. *International journal of evidence based coaching and mentoring*, 6(1):32–42.

[Stolcke et al.2000] 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.

[Weizenbaum1966] Joseph Weizenbaum. 1966. Eliza - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, January.

[Wexley and Baldwin1986] Kenneth N Wexley and Timothy T Baldwin. 1986. Posttraining strategies for facilitating positive transfer: An empirical exploration. *Academy of Management Journal*, 29(3):503–520.

[Zhou and Oldham2001] Jing Zhou and Greg R Oldham. 2001. Enhancing creative performance: Effects of expected developmental assessment strategies and creative personality. *The Journal of Creative Behavior*, 35(3):151–167.

If you repeat your interlocutor's syntactic structure, you are likely to repeat her pronunciation, too

Marie Postma-Nilsenová
TiCC, School of Humanities
Tilburg, The Netherlands
M.Nilsenova@uvt.nl

Lisette Mol
TiCC, School of Humanities
Tilburg, The Netherlands
L.Mol@uvt.edu

Naomi Kamoen
TiCC, School of Humanities
Tilburg, The Netherlands
N.Kamoen@uvt.nl

Abstract

Past studies showed that in dialogue, interlocutors copy each others' behavior in various ways. Copying on one grammatical level leads to increased chances of copying on other levels as well, a so-called 'alignment boost'. The identification of specific alignment boosts offers important insights into the architecture of language comprehension and production because it highlights relations between different types of linguistic representations. We examine the possibility of a direct influence from grammar on sound with no involvement of the conceptual system. In priming experiments with non-words, we show the existence of a direct syntactic alignment boost to segmental phonology. The outcomes are of relevance to models of language processing.

1 Introduction

Past research has established beyond doubt that conversation participants frequently take over each others' structural and phonetic/phonological choices. One of the first investigations describing *syntactic* imitation was Schenkein's (1980) analysis of repetitions in burglar conversations over walkie-talkies. The phenomenon has been observed many times since, both in experimental studies and in studies of natural interactions (for an overview, see Pickering & Ferreira, 2008). Also in the area of *phonetics* and *phonology*, repetition of recently produced or perceived patterns has been widely documented, e.g., for the pronunciation of vowels and consonants, pitch accent, speech rate, and low and high boundary tones (Natale, 1975; Gregory & Hoyt, 1982; Giles, Coupland & Coupland, 1991; Pardo, 2006; Delvaux & Soquet, 2007; Nilsenová, Swerts, Houtepen & Dittrich, 2008). In the newly proposed forward model by Pickering and Garrod (2013), repetition starts during the process of language comprehension in the form of covert imitative behavior that helps the perceiver predict upcoming linguistic representations. Comprehension and production are not isolated processes, rather, they are interweaved through imitation.

The tendency to "reuse what has been used" offers rich testing grounds for theories of language architecture for two reasons: First of all, if certain representations - e.g., abstract syntactic representations that are independent of meaning and sound (Pickering & Ferreira, 2008) - are reused, it means that they form a part of the speech planning process. Second, it has been observed that the repetition can be enhanced if other representations are repeated as well, for instance, syntactic imitation gets a 'boost' from a repetition of the head verb (Branigan, Pickering, & Cleland, 2000; Branigan, Pickering, McLean & Cleland, 2007). The occurrence of such a boost has been interpreted as evidence that some, but not all, levels of representations are related in the sense of percolating activation (Pickering & Garrod, 2004), see Figure 1.

In the context of the theoretical discussion regarding the links between linguistic representations and their reuse in dialogue, the study reported here has two objectives. First, we set out to replicate in another language the results of Branigan, Pickering, and Cleland (2000) who observed a boost from repetition of a head verb on syntactic imitation. Second, we test the possibility of a syntactic boost on phonology while excluding the involvement of the lexicon, making use of two different experimental paradigms: a verb invention task and a rhyming task. To our knowledge, the (direct) effect of syntax on phonology has not been examined in the context of alignment studies.

The idea that syntax determines phonological operations has been around for some time, e.g., according to Bierwisch (1966), syntax 'feeds phonology' when syntactic output is converted into phonological output. The relation can also be illustrated with phenomena such as liaison, syntactically determined segmental duration or accent placement (Klatt, 1975; Selkirk, 1974). However, most linguistic studies appear to use suprasegmental phenomena for the argument that syntax drives phonology, e.g., segment duration and coarticulation can presumably be included under prosody given that they are related to prosodic boundaries. It could, in fact, be the case that various types of phonological segments are differently affected by the speaker's syntactic choices (Santesteban, Pickering, and McLean, 2010), nonetheless, the claim made by Pickering & Garrod's model of language processing is stronger because it presumably concerns phonology as a whole. Therefore, we expect a possible boost of syntactic repetition to occur on a segmental level as well. Any other outcome would suggest that the levels of linguistic representations postulated in the Alignment model need to be refined and, possibly, their relations to other representations revised.

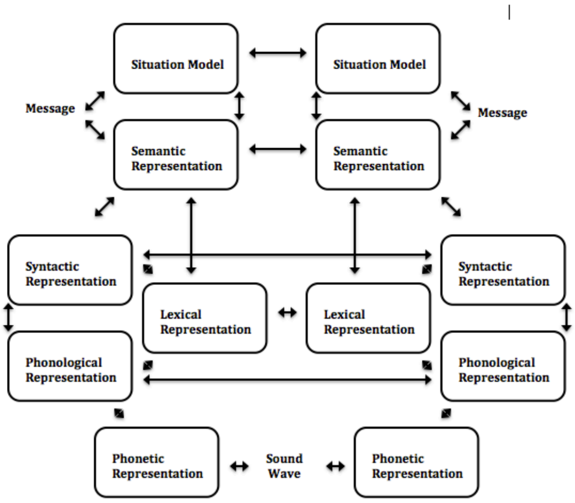


Figure 1: The Interactive Alignment Model (reproduced on the basis of Pickering & Garrod, 2004:177).

2 Current study

Below, we report the experimental methods and results of three experiments designed to test the existence of a lexical boost on syntax (Experiment 1), a syntactic boost on phoneme selection in a task involving invented verbs (Experiment 2) and a syntactic boost on rhyme pronunciation with stimuli containing nonexistent brand names (Experiment 3). In the first experiment, we sought to replicate the results of Branigan, Pickering, and Cleland (2000) for English by adapting their experimental design for Dutch. In the second and third experiment, we made use of two different experimental methods to test the link between syntax and phonology.

To analyze the data, we made use of multilevel models. In traditional ANOVAs the variance due to items and the variance due to respondents cannot be estimated simultaneously. As a consequence, the total variance is underestimated, which causes $H0$ to be rejected too easily (see Quené & Van den Bergh, 2004). To decrease the risk of type 1 errors, we applied multilevel models in the current study. Such models do allow for estimating the between item variance and the between respondent variance simultaneously. For example, in the multilevel model for experiment 1, the percentage of alignment is estimated for the boost and no boost conditions. In addition, the model allows these means to vary between items (one item may elicit more alignment than another) and between respondents (one respondent may align more frequently than another). These variances are estimated simultaneously, so in fact a cross-classified model is in operation (see Quené & Van den Bergh, 2004). The alignment percentages for the boost and no boost conditions can be compared in a contrast test (Bosker & Snijders, 1999; Goldstein, 2003), which yields a χ^2-distributed test statistic. For a formalization of this model and further explanation, we refer to Appendix 1.

3 Experiment 1

In the first experiment, we made use of the method originally due to Branigan, Pickering, and Cleland (2000), in order to replicate their study of English verb-repetition boosts on syntactic priming.

3.1 Method

Participants

Twenty-two Dutch speakers (14 female; mean age 18;8) were recruited from a Dutch University student population and received course credits for their participation.

Design

The participants were randomly assigned to two experimental conditions, with or without lexical boost (Boost and No Boost, respectively).

Materials

The participants took part in a confederate-governed task of describing 28 drawings (12 ditransitive stimuli (see Figure 2) + 16 monotransitive fillers (see Figure 3)), while being primed alternatively with a syntactic structure of the form 'ditransitive verb + direct object + prepositional indirect object (e.g., "The pirate is giving the book to the captain.") and a structure of the form 'ditransitive verb + (non-prepositional) indirect object + direct object (e.g., "The pirate is giving the captain a book."). For their description, they were asked to use the verb given under the drawing. In the condition with lexical boost, the verb was the same as the verb used in the confederates prime, in the condition without lexical boost, the verbs differed.

To balance for order effects and verb effects, in both conditions, there were 4 confederate variants with structures alternating per verb. The sentences we used were Dutch translations of the sentences employed by Branigan et al. (2000) in their picture-matching task. An experimental pilot (N=33) revealed a possible effect of the monotransitive fillers on the experimental trials; in Dutch, unlike in English, ditransitive verbs such as *geven* 'to give' or *overhandigen* 'to hand over' can be used in monotransitive constructions as well. Therefore, we adapted the fillers in such a way that they resembled the experimental trials in terms of length and syntactic complexity by including a propositional phrase (e.g., 'The boy is drawing a picture on the board', instead of the original 'The boy is drawing a picture').

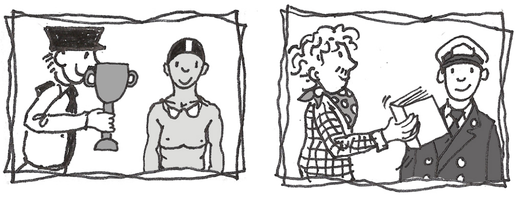


Figure 2: Examples of experimental stimuli depicting ditransitive events.

Procedure

During the experimental session, the participant was seated opposite to the confederate who pretended to be a participant as well. The experimental leader was present in the same room to answer questions and make sure that the participant followed the experimental instructions. The experiment was presented as a game of describing and finding pictures, where both the correctness of the response (picture found) and the time needed to do so would be compared across conditions. The participants were explicitly told that rather than performing the task quickly, they should attempt to be as precise as possible. The output for both conditions was recorded on paper by the confederate, and the dialogue was digitally recorded with the help of MacBook computer by the experimenter. After each experimental session, the transcripts were compared to the audio recordings and corrected if necessary.

The confederate and the participant were taking turns in describing the pictures (see Figure 2), with the confederate always initiating the turn (in other words, priming the participant). The confederate picture set included full sentence descriptions of the pictures but in order to maintain the appearance of being a participant as well, the confederate pretended to be making up the descriptions on the spot. The participant was not aware of what was in the confederate set, but assumed that it resembled his/her own.

After the experimental session, the experimental leader asked both the confederate and the participant if they noticed anything unusual. Only after that did she disclose the real purpose of the experiment and the role of the confederate.

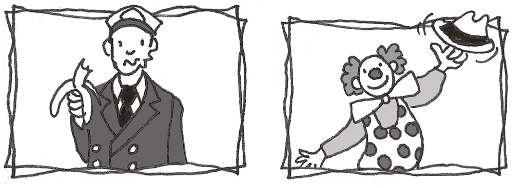


Figure 3: Examples of experimental fillers used in the confederate task in all three experiments.

Scoring

The trials in which the participant used the same syntactic construction as the confederate were scored as 1 and the trials where the participant used a different construction, be it the alternative ditransitive structure or a monotransitive one, were scored as 0.

Results

Table 1 presents the average probability of alignment for the Boost and No Boost conditions. Results show that participants aligned more often in the Boost condition than in the No Boost condition ($\chi^2 = 12.25$; df = 1; $p < 0.001$).[1] As the systematic between-person variance is estimated to be zero (see Table 1), the difference between the Boost and No Boost conditions is large as compared to the systematic differences between respondents. In comparison to the systematic differences between items, the size of the effect can be classified as medium (Cohen's $d = 0.41$).

Condition	**Mean**	$\mathbf{S}^2{}_{items}$	$\mathbf{S}^2{}_{persons}$
Boost	.76 (1.14)***	1.14	0
No Boost	.55 (1.55)		0

Table 1: Parameter estimates of imitation probability for the boost and no-boost condition.
Note. *** $p < .001$. For the sake of convenience, the mean alignment probabilities provided in proportions and in the Logits used for the analysis (between brackets). The variances are only provided in Logits. The item variance is estimated once, for the Boost and No Boost conditions together.

Discussion

The first experiment showed that syntactic priming received a lexical boost in the condition in which participants were using the same verb as the confederate in his/her prime (the boost condition). The result is a replication of the finding reported by Branigan, et al. (2000) for English.

4 Experiment 2

In the second experiment, we explored the effect of a syntactic boost on phonological alignment. In order to test for the relationship directly, we aimed to exclude the effects of the lexicon that is likely to facilitate phonological alignment in spontaneous data.

4.1 Method

Participants

Twenty-four speakers of Dutch (15 female; mean age 19;3) drawn from the same participant population as in Experiment 1 took part in the experiment. None of the speakers took part in the other two experiments in this study.

Design

The participants were randomly assigned to two experimental conditions, syntactic Boost and No Boost.

Materials

Participants were filling in an invented verb into a blank of the form 'NP who IO DO' (e.g., *De man die de non een appel…* "The man who…the nun an apple.") or 'NP who DO PO' (e.g., *De man die een appel aan de non…* - "The man who … an apple to the nun"). The systematic variation in the confederates verbs consisted (1) in the number of syllables (one or two) and the initial phoneme (a vowel or a consonant), see Table 2. In total, there were 24 experimental trials.

Initial Phoneme	**Monosyllabic**	**Disyllabic**
Vowel	*oeft*	*oegert*
	aapt	*eivelt*
	oot	*affelt*
	iert	*uitert*
	eift	*iemelt*
	eemt	*okkelt*
Consonant	*proest*	*manilt*
	kniert	*pippelt*
	bort	*lippert*
	vlaapt	*zachelt*
	slinkt	*poenkert*
	loept	*niesert*

Table 2: Monosyllabic and disyllabic nonwords used as primes in Experiment 2.

Procedure

Same as in Experiment 1.

Scoring

The nonsense verbs created by the participants were transcribed by the experimental leader dur-

---

[1] The size of an effect is often classified in relation to the standard deviation (Cohen 1988). In the kind of multi-level models that we have used, different sources of variance are modeled and so different measures for the effect size may be calculated (i.e. the effect size relative to the between-item standard deviation and the effect size relative to the between-person standard deviation).

ing the experimental session, as well as digitally recorded. The transcriptions were made in such a way as to reflect the rules of the Dutch spelling system and checked against the audio recordings first by the experimental leader and subsequently by another linguist. The confederate's and participants' nonwords were first transcribed in IPA by an independent condition-blind linguist in accordance with the mainstream Dutch phonological system (Appel et al., 2001).

We calculated the proportion of phonological alignment by comparing broad and narrow phonological transcriptions of the prime and the target. For the broad phonological comparison, we scored responses as 1 if there was at least one phoneme in the prime and in the participants' response that had an identical manner or place of articulation as the prime (again disregarding its position and excluding the $3rd$ person singular morpheme), and 0 otherwise. For the narrow phonological comparison, we scored the responses as 1 if at least one identical phoneme in the prime and in the participants' response was present (disregarding its position and the $3rd$ person singular $-t$ at the end which was present in all responses); otherwise, the response was coded as 0.

Results

Table 3 shows for the Boost and No Boost conditions the percentage of alignment, both for the broad phonological scoring system and the narrow phonological scoring system. When a broad phonological scoring system is used, we find that participants align equally often in the Boost condition and the No Boost condition (χ^2 = 0.05; df = 1; p = 0.82). However, when a narrow phonological analysis of alignment is performed, differences can be observed: participants align more often in the Boost condition than in the No Boost condition (χ^2 = 6.25; df = 1; p = 0.01; Cohen's d = 0.31).[2]

Discussion

The results of the second experiment indicate that there is a link between the syntactic and the phonological component that does not have to be mediated by the lexicon. In particular, when speakers repeat the syntactic choices of their dialogue partner, they are also more likely to align phonologically. The phonological adaptation, however, is rather subtle and, at least in this experiment, was only revealed when a narrow phonological transcription was used to score the participants' responses.

A generalization of the outcome of the experiment might be difficult, given the low ecological validity of the task used in the experimental procedure. Therefore, we conducted a third experiment in which we again tested the presence of a syntactic boost on phonology with a different task involving the pronunciation of unknown brand names for products depicted on the drawings used in the previous experiments.

Set	Type	Brand	Set	Type	Brand
1	filler	*Novita*	7	filler	*Teps*
	filler	*Mapri*		filler	*Fobat*
	prime	*Slent*		prime	*Prievais*
	target	*Flant*		target	*Crevi's*
2	filler	*Zilko*	8	filler	*Savin*
	filler	*Altreno*		filler	*Gikmer*
	prime	*Kedélen*		prime	*Vaik*
	target	*Hedelen*		target	*Tike*
3	filler	*Blarkin*	9	filler	*Ritrabo*
	filler	*Xepon*		filler	*Hijntes*
	prime	*Xail*		prime	*Nóreo*
	target	*Rile*		target	*Toreo*
4	filler	*Walmits*	10	filler	*Quotrepi*
	filler	*Crendum*		filler	*Jovent*
	prime	*Drend*		prime	*Bingles*
	target	*Grand*		target	*Fringles*
5	filler	*Volstar*	11	filler	*Pladow*
	filler	*Dapens*		filler	*Krepo*
	prime	*Metálogis*		prime	*Welsprie*
	target	*Protalogis*		target	*Depsprit*
6	filler	*Unalem*	12	filler	*Obitan*
	filler	*Elanit*		filler	*Lantadi*
	prime	*Zappel*		prime	*Njugels*
	target	*Qappel*		target	*Kugels*

Table 4: Brand names used as primes and fillers in Experiment 3.

5 Experiment 3

In the third experiment, we examined the direct relationship between syntax and phonology (with no intervention of the lexicon) with the help of a 'rhyming task' implemented in a design akin to the previous two experiments.

---

[2] An analysis of alignment on the number of syllables and the initial phoneme (vowel or consonant) showed no significant effects of the experimental manipulation.

Scoring	Condition	Mean	S^2_{items}	$S^2_{persons}$
Broad Phonetic	Boost	.78(1.24)	1.62	0
	No Boost	.77 (1.22)		0.14
Narrow Phonetic	Boost	.49(-0.04)*	1.89	0
	No Boost	.39 (-0.46)		0

Table 3: Parameter estimates of imitation probability for the boost and no-boost condition.
Note. * $p < .05$. For the sake of convenience, the mean alignment probabilities provided in proportions and in the Logits used for the analysis (between brackets). The variances are only provided in Logits. The item variance is estimated once, for the Boost and No Boost conditions together.

5.1 Method

Participants

Forty speakers of Dutch (20 female; mean age 21;7) drawn from the same participant population as in Experiment 1 and 2 took part in the experiment. None of the speakers took part in the other two experiments in this study.

Design

The participants were randomly assigned to two experimental conditions (with and without syntactic boost). The dependent variable was phonological alignment ("rhyming").

Materials

We made use of the same drawings depicting ditransitive events as in experiment 1 with the addition of a nonword "brand-name" before each object in the sentence, with 12 experimental trials and 12 fillers (see Table 4), with two different order variations. To prevent lexical priming, the head verbs used in the prime-target pairs were always non-identical. Prior to the experiment, a pretest was conducted with a different group of participants from the same population (N = 33). The goal of the pretest was to determine the preferred pronunciation of the invented brand names. In the primes used by the confederates in the subsequent experiment, we only used the non-preferred pronunciation.

Procedure

Same as in Experiment 1 and 2. In the instructions given to the participants we asked them to read the sentences under the drawings to indicate which drawing their partner should search (e.g., "The teacher is handing over the *Slent* banana to the swimmer.").

Scoring

Participants' pronunciation of the invented brand names was transcribed by two research assistants blind to condition. The responses were scored as 1 if the pronunciation rhymed with the pronunciation of the confederate and 0 otherwise.

Results

Table 5 shows the average probability of alignment for the boost and no boost conditions. Results show that participants align more often under boost conditions than under no boost conditions (χ^2 = 7.71; df = 1; $p < 0.01$; Cohen's d = 0.57).

Condition	Mean	S^2_{items}	$S^2_{persons}$
Boost	.40 (-0.39)*	0.89	0
No Boost	.28 (-0.93)		0

Table 5: Parameter estimates of imitation probability for the Boost and No Boost condition.
Note. * $p < .05$. For the sake of convenience, the mean alignment probabilities provided in proportions and in the Logits used for the analysis (between brackets). The variances are only provided in Logits. The item variance is estimated once, for the Boost and No Boost conditions together.

Discussion

The third experiment confirmed the partial finding of Experiment 2. We found that participants were more likely to use the dispreferred pronunciation of an unknown brand name if they were repeating the same syntactic structure as the confederate and the pronunciation rhymed with the confederate's immediately preceding choice.

6 General Discussion

The series of experiments reported here focused on the relations between two pairs of linguistic representations, the lexicon and syntax (Experiment 1) and syntax and phonology (Experiment 2 and 3). These relations were examined in the context of repetitions in an interactive game with alignment boosts. Earlier studies of the link between syntax and phonology, in particular the existence of *phonological* boost on syntax, offer inconclusive results. In a hallmark study, Bock (1986) reported that prime words that were semantically related to entities represented visually gave rise to active/passive constructions in which the semantically related words came first. This finding, however, was not replicated for primes that were phonologically related to words describing entities in the visual material (e.g., the prime frightening did not give rise to constructions starting with lightning). Bocks conclusion was that unlike semantics, phonology did not influence syntactic formulation. Similarly, Cleland and Pickering (2003) found no enhanced priming effect of phonological similarity on noun-phrase structure (a complex noun phrase containing a relative clause vs. a simple noun phrase). In their study, this result was contrasted with the enhanced priming effect of semantically related nouns. Again, it was taken to suggest that phonology does not appear to give a boost to syntactic alignment. A more recent study of between-language syntactic priming in constructions involving cognates, though, suggested that phonology may affect syntax at least to some extent (Bernolet, Hartsuiker, & Pickering, 2012). In particular, in a study with Dutch-English bilinguals, cognates boosted syntactic priming, while non-cognates did not. This result seems to be in line with an earlier observation by Lee and Gibbons (2007) that the preference for metrical structure (the rhythmic alternation between stressed and unstressed syllables) affects the (syntactic) decision to use a complementizer. It is also in line with the outcome of Santesteban, Pickering, and McLeans (2010) experiment showing that the use of semantically unrelated homophones boosts syntactic priming; in their experiments the effect was as strong as the effect of a lexical boost.

In sum, it appears that with respect to the direction from phonology to syntax, the relation between syntax and phonology might be more complex than the representation currently included in the Alignment model. The outcomes of the experiments reported here indicate that a similar conclusion might be drawn for the relation from phonology to syntax. Future research needs to disentangle how various types of phonological representations (segmental/suprasegmental, word-initial, accented, etc.) affect and are affected by syntactic repetition, for instance by measuring the effects of syntactic boost on accent placement.

7 Conclusion

Participants in an interactive task repeated the linguistic choices of their partners more often if they were instructed to repeat the same head verb (Experiment 1) or the same syntactic structure (Experiment 2 and 3) in the same sentences. The outcomes of the experiment suggest the existence of alignment boosts from the lexicon to syntax and from syntax to phonology. The second type of boost appears to affect various phonological segments to a different degree, which suggests that the levels of representations currently represented in the Alignment model need to be refined.

Acknowledgments

The authors would like to thank Marije van Amelsvoort, Annabel Verkerk, Floor de Vries and Annika van Heeswijk for their help with collecting data for the experiments, and to Holly Branigan and Huub van den Bergh for methodological and statistical comments, respectively. Experiment 1 and 2 have earlier been presented at *CogSci* 2010 with a different type of scoring for Experiment 2 and a different statistical analysis.

References

Sarah Bernolet, Robert J. Hartsuiker and Martin J. Pickering. 2009. Persistence of emphasis in language production: A cross-linguistic approach. *Cognition*, 112: 300–317.

Manfred Bierwisch. 1966. Regeln fuer die Intonation deutscher Saetze. *Untersuchungen Ueber Akzent und Intonation im deutschen, Studio Grammatica*, pp. 99–201. Berlin: Akademie Verlag.

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18: 355–387.

Roel J. Bosker and T.O.M. Snijders. 1999. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. New York.

Holly P. Branigan, Martin J. Pickering and Alexandra A. Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition,* 75:B13–B25.

Holly P. Branigan, Martin J. Pickering, Janet F. McLean and Alexandra A. Cleland. 2007. Participant role and syntactic alignment in dialogue. *Cognition*, 104: 163-197.

Alexandra A. Cleland and Martin J. Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49:214–230.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Veronique Delvaux and Alain Soquet. 2007. The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, 64:145–173.

Howard Giles, Nikolas Coupland and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland & N. Coupland (Eds.) *Contexts of Accommodation*, pp. 1–68.

Harvey Goldstein. 2003. *Multilevel statistical models*. London: Edward Arnold.

Stanford W. Gregory and Brian R. Hoyt. 1982. Conversation Partner Mutual Adaptation as Demonstrated by Fourier Series Analysis. *Journal of Psycholinguistic Research*, 11:35–46.

Ming-Wei Lee and Julie Gibbons. 2007. Rhythmic alternation and the optional complementiser in English: New evidence of phonological influence on grammatical encoding. *Cognition,* 105:446–456.

Michael Natale. 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology,* 32:790–804.

Marie Nilsenová, Marc Swerts, Veronique Houtepen and Helene Dittrich. 2009. Pitch adaptation in different age groups: boundary tones versus global pitch. *Proceedings of Interspeech,* September 6-10, Brighton.

Jennifer S. Pardo. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America,* 119:2382–2393.

Martin J. Pickering and Victor S. Ferreira. 2008. Structural Priming: A Critical Review. *Psychological Bulletin,* 134(3):427–459.

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

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences,* 36:329–392.

Hugo Quené and Huub Van den Bergh. 2004. On multilevel modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43: 103-121.

Hugo Quené and Huub Van den Bergh. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59: 413-425.

Mikel Santesteban, Martin J. Pickering and Janet F. McLean. 2010. Lexical and phonological effects on syntactic processing: Evidence from syntactic priming. *Journal of Memory and Language*, 63: 347-366.

J. Schenkein. 1980. A taxonomy for repeating action sequences in natural conversation. In B. Butterworth (Ed.). *Language production*, Vol. 1, pp. 21–47. London: Academic Press.

Elisabeth Selkirk. 1974. French liaison and the X notation. *Linguistic Inquiry*, 573–590.

Appendix: Multi-level models

As an example, we will elaborate on the multi-level model applied in experiment 1. In this model, the occurrence of alignment is estimated separately for the boost and no boost conditions. This is done in Logits, because the estimations concern a binomial dependent variable. In addition, between-item variance and between-person variance are allowed.

Equation A1 gives a formalization of the model applied in experiment 1. In this model, $Y_{(jk)}$ indicates whether or not participant j (j = 1, 2,. . . , 28) aligns with the confederate for item k (k = 1, 2,. . . ,14). In addition, there are two dummies (D), one for the boost conditions (D_BOOST$_{(jk)}$), and one for the no boost conditions (D_BOOST$_{(jk)}$). These dummies can be turned on if the observation matches the prescribed type. Using these dummies, two probabilities are estimated, representing the occurrence of alignment under the boost and no boost conditions (β_1 and β_2). These may vary between items (v_{0k}) and between persons (u_{1j0}, u_{2j0}).

Equation A1:

$$\text{Logit } (\text{Y}_{(jk)}) = \text{D\_BOOST}_{(jk)} \; (\beta_1 + u_{1j0}) + \text{D\_NOBOOST}_{(jk)} \; (\beta_2 + u_{2j0}) + v_{0k}.$$

The model in Equation A1 can be described as a cross-classified model (Quené & Van den Bergh, 2008), as the model accounts for each observation to be nested within items and persons at the same time. All residuals are normally distributed with an expected value of zero, and a variance of respectively $\text{S}^2{}_{u1j0}$, u_{2j0}, and $\text{S}^2{}_{v0k}$. Please note that in this model the item variance $\text{S}^2{}_{v0k}$ is estimated only once for boost and no boost conditions together. This is a constraint of the model.

Learning new words in unfamiliar frames from direct and indirect teaching

Katharina J. Rohlfing
CITEC, Emergentist Semantics
Bielefeld University
`kjr@uni-bielefeld.de`

Juana Salas Poblete
CoR-Lab
Bielefeld University
`juana.salas@uni-bielefeld.de`

Frank Joublin
Honda Research Institute
`frank.joublin@honda-ri.de`

Abstract

In our study, we aimed at investigating how two years old children make use of the pragmatics in order to learn new words from an ongoing interaction. We operationalized the situational pragmatics by frames as introduced by developmental psychologists. The basic logic was to place 30 children at the age of 27 months in situations, in which they can barely rely on their prior pragmatic knowledge. Instead, they depend on extracting more information from the observed interaction. Our hypothesis was that when learning in unfamiliar frames, children in the indirect teaching condition would take advantage of the modeled behavior to identify with one of the communicative partners and thus to learn new words.

1 Introduction

Imagine a family eating breakfast. The children have invited a friend over, who asks for a napkin: “Can I have a Zewa?”. In German, “Zewa” is an eponym standing for piece of a kitchen roll or a paper towel. She is corrected by the other children who inform her that at their home, she must ask “Can you give me a napkin?” Both sentences are syntactically and semantically different but on a pragmatic level, they lead to the same goal. And obviously, one must know how to frame such a goal (i.e. which verbal action to choose) to successfully achieve it. In this work, we were interested how children learn the pragmatic frame as a form of an appropriate action.

The concept of frames was introduced to developmental psychology by Bruner (1983) as an implicitly encoded social behavioral pattern acquired through experiencing social interactions in one’s cultural environment (Bruner, 1983; Fogel, 1993; Tomasello, 1999; 2003). Frames are supposed to give children access to the principles that guide social interaction as they provide “predictable, recurrent interactive structures” (Ninio & Snow, 1996, p. 171) that scaffold the child’s emerging understanding of new linguistic labels (Tomasello, 2003). In this sense, embedding a new word within a familiar frame results in the reduction of the information load on the child as this word will be perceived within a familiar routine and “the process of word learning is constrained by the child’s general understanding of what is going on in the social situation in which she hears a new word” (Tomasello & Akhtar, 200: 182).

The importance of frames for learning has been acknowledged by Fogel (1993), Ninio & Snow (1996) and Tomasello (2003). However, pragmatic knowledge is difficult to investigate as it provides action frames within which a successful interaction takes place and therefore can usually be observed only implicitly. To our knowledge, ours is the first experimental study actively manipulating the pragmatic frame in the context of word learning to explore its influence on learning success. By pragmatic frame, we understand an interaction protocol involving actions in a sequence that is coordinated with the interaction partner. The coordination evolves as a routine: Performing a speech act such as labeling a new object, a competent speaker knows that this goal has to be framed by (a) looking at the other person, (b) pointing in the direction of an object and (c) uttering a label (see Figure 1, column “familiar frame”). In this familiar routine, (c) can be perceived as a slot, within which new information is provided and can be easily picked up.

To date, investigation of pragmatic frames concentrated on whether and at which age children master a particular routine. E.g., Franco & Butterworth (1996) have shown that at the age of 16 months, children learn to visually check whether the interlocutor is attentive before they actually point to something, which is basically the part (b) in the labeling speech act described above. Another strand of research is devoted to the link between language acquisition and imitation skills drawing from the fact that pragmatic frames consist of an appropriate action. Thus, it is likely that children acquire such frames through their imitation skills. Interestingly, a strong link between imitation capabilities and language learning is assumed suggesting not only that such frames might be a form of cultural transmission but also that children need to learn to apply it in a reverse role (Tomasello, 1999). Studies found that children's ability to imitate in a reverse role was related to various measures of language acquisition for 18 month olds (Carpenter et al., 2005; Herold & Akhtar, 2008). Thus, children need to imitate in a reverse role "to learn to use bidirectional communicative symbols" (Carpenter et al., 2005, p. 275).

However, the investigation focuses mostly on direct teaching scenarios. Yet, there is increasing evidence suggesting that indirect teaching scenarios might be even more fruitful learning environments in conveying skills that are related to pragmatic knowledge. This evidence is coming from a variety of different research strands. One strand is dedicated to overhearing studies. In these studies, children are not addressed directly but rather hear the tutor talking to another person and pick up a learning content from this indirect teaching (Akhtar, 2005; Akhtar, Jipson, & Callanan, 2001; Floor & Akhtar, 2006; Gampe et al., 2012). In theses studies, it has been found that when a reciprocal social interaction is guaranteed, young children learn words similarly well as in direct teaching scenarios. The other strand of research is coming from sociolinguistic studies emphasizing that in many cultures, children are taught how to behave and act appropriately within the community's interaction rules (Heath, 1983; Ochs, 1986; Pye, 1986; Schieffelin, 1986; Scollon & Scollon, 1981), therefore centering learning processes in the field of pragmatics.

Yet another strand of research comes from work by Oshima-Takane and colleagues (1996), who demonstrated that children with more multiparty interaction experience had better success in the acquisition of personal pronouns. This line of investigation pursues the idea that not only the acquisition of the lexical item itself takes longer in children who are less experienced with multiparty interactions, but their lack of opportunity to observe its correct usage – i.e. the unfamiliarity with contexts in which personal pronouns are typically used – delays production.

All together, the various strands of research speak to the possibility that the acquisition of pragmatic frames is particularly facilitated in polyadic interactions. However, to date, neither the question of how pragmatic frames are acquired nor in which learning environment they might be learned has been addressed in word learning studies. This paucity is due to the fact that pragmatic knowledge is implicit to the process of language acquisition: Children make use of culturally established routines and it is difficult to design a new interaction protocol consisting of truly new actions. Thus, we think that both, (a) a defined routine consisting of a fix interaction protocol and (b) new actions within it are required to appropriately test the acquisition of pragmatic frames.

2 Designing unfamiliar frames

When investigating the acquisition of pragmatic frames, it is necessary to ensure that children bring little prior knowledge of action into the testing situation. More specifically, in the study by Gampe and colleagues (2012), it was tested whether eighteen month-old children will learn new words from overhearing, even though the frames that were used to introduce the new words were not established as a labeling routine. A labeling routine would be to say, "look, this is a toma!", but in Gampe et al. (2012, p. 5), the experimenter said "I'm going to show you the toma. Do you want to see the toma?". Thus, basically, a 'showing' frame was used to introduce the novel label, which is definitely not a typical labeling routine but nonetheless a familiar frame. From the results in this study, it was concluded that children could learn a novel label even in less transparent situations, in which not a typical labeling frame was used. With respect to the pragmatic skills, it is interesting to see that the use of (almost) any kind of pragmatic frame will facilitate learning of words in children. However, the question of how such frames are established remains barely investigated. As already mentioned above, it is difficult to create truly new actions, i.e. actions that the children have to learn without drawing advantage on their prior knowledge. In our attempt to solve this problem, we created a frame with unfamiliar elements in an interaction protocol – on the basis of Pepperberg’s model/rival labeling routine (1997; 2002) – as a condition in which children needed to learn a new behavior to be able to participate appropriately in the interaction. Based on Bandura’s work (1971), Pepperberg (2002) developed this routine for a grey parrot acquiring labels for fifty objects, seven colors, number labels up to eight, categories, etc. The linguistic abilities of the grey parrot trained with the model/rival technique exceeded simple naming of individual items as he was able to combine these labels and use them referentially, which enabled him to identify, classify, request or decline over a hundred items. On a pragmatic level, he was able to distinguish simple speech acts and communicative roles (Pepperberg, 1992). Pepperberg (2002) argues that the model/rival technique maximizes the level of explicitness in presenting reference, functionality and nonverbal context framing the social interaction: During teaching sessions, the parrot observed a dialog taking place between two experimenters. One of the experimenters acted as tutor and the other as both model and the parrot’s rival for the tutor’s attention. The dialog consisted of a fixed question-answer-routine: The tutor asked for the denomination of an object and the model/rival gave either a correct or an incorrect answer. This in turn triggered either a positive, reinforcing feedback or a negative, corrective feedback. The positive feedback consisted of verbal praise and the possibility for the human (and later the parrot, were it correct) to play with the object — which was the ultimate goal. The negative feedback consisted of a verbal scolding, interruption of eye contact and retraction of the object. Tutor and model constantly changed roles so the parrot learned to separate the role from the person.

How this method from an animal study can be usefully applied in studies with children was shown in Pepperberg and Sherman (2002). The underlying argument was that children with special needs might benefit from the model/rival technique: Instead of requiring the child to react to parts of an interaction (e.g. a question), the behavior modeled in an indirect teaching scenario was assumed to demonstrate the appropriate verbal and nonverbal behavior in a holistic way. Pepperberg and Sherman (2002) tested the model/rival paradigm with 24 children with various disabilities: autism, physical disabilities with developmental delays, and attention deficit hyperactivity disorder. The rewards applied in model/rival training were modified for children: Instead of physical objects, the children received the opportunity to interact with the tutor and the model by singing a song or playing a game. All children had received conventional one-to-one treatment before the study but without obtaining an important improvement in their condition. With the model/rival training, however, all children made improvements in their interactive communicative skills even though this study did not primarily focus on the acquisition of new word knowledge but on the acquisition of appropriate behavioral patterns. This provides strong support to the idea that – in contrast to direct teaching – indirect teaching seems to facilitate learning under certain conditions, but still systematic application in the field of language acquisition is lacking.

Motivated by these findings, we aimed to apply this technique to language acquisition with typically developed children to evaluate the effects of indirect teaching with respect to learning pragmatic frames. It can be argued that the model/rival paradigm is similar to the so called overhearing scenarios (Akhtar, 2005; Akhtar, Jipson, & Callanan, 2001; Floor & Akhtar, 2006). There are, however, some crucial differences in these two scenarios. While in overhearing scenarios children are not a part of an interaction, in our scenario, children were positioned as onlookers to an instructive dialog between two adults. In addition, not only did the children hear a new word introduced – as in overhearing scenarios – but they were also presented with a model of a holistic verbal and nonverbal behavior. Thus, our setting can be considered a very specific form of an overhearing scenario.

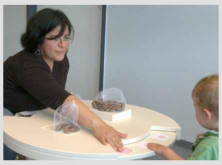
	Familiar frame	**Unfamiliar frame**
		
Ostension	eye contact + calling the child by her given name	
Highlighting	pointing	illuminating/elevating
Question	“What’s this?” for nouns; “What’s the color of this block?” for color adjectives; “How many buttons are these?” for number words	
Answer	word production	nonverbal behavior: touching the correct display

Figure 1: Unfamiliar frame in a comparison to a familiar frame.

As can be seen in Figure 1, in our scenario, we manipulated two parts of a typical question-answer routine: The highlighting of the object or its properties and the way the children had to provide their answer.

In various studies, it has been shown that socio-pragmatic cues such as eye-gaze, pointing, touching or manipulation (see summary in Booth et al., 2008) can draw children's attention to an object or its properties. However, for younger children, it has been shown that before children begin to regard the socio-pragmatic cues within an interaction, their attention can be guided by perceptual properties of the objects themselves. For example, when hearing an object labeled, 10 months old will associate this label with a more salient object (Pruden et al., 2006). Thus, designing the unfamiliar pragmatic frame, we make use of the fact that children's attention to an object can be directed not in a familiar way (by pointing) but in an unfamiliar way by lighting up the object's location or elevating it mechanically. This specific way of highlighting was made possible by a table that was designed for this study (see Figure 2).

In addition to the way of highlighting an object, the children's answers in this interaction were also designed in an unfamiliar way. We reasoned that almost any action that elicits a word production from a child is familiar. Thus, we rather requested a nonverbal behavior from the child in form of placing the hand on the one of three displays in front of them. The object and the displays are depicted in Figure 3 below.

Based on the above referred sociolinguistic and laboratory studies about learning in multi-party contexts, for our study, we assumed that the "benefit [of multi-party learning] involve pragmatic skills rather than the more strictly linguistic skills such as vocabulary size" (Barton & Tomasello, 1991, p. 518). Therefore, the research question was whether children acquire pragmatic knowledge better in direct or indirect teaching conditions. Although a certain agreement exists in developmental pragmatics that frames play a role in language acquisition, this role has been claimed only for direct teaching interactions. Thus, our study fills a gap as it compares direct and indirect teaching scenarios with respect to how pragmatic frames are acquired and whether multi-party interactions can contribute to it.

We hypothesized that children in the indirect teaching condition would score significantly better than children taught directly by taking advantage of the presence of a model, thereby facilitating imitation of the involved appropriate action.

3 Method

In this experiment, in addition to our data obtained during the interaction between the child and the experimenter(s), we asked the accompanied parents to fill out two questionnaires: The short version of the ELFRA-2 (Grimm & Doil, 2006) – a German equivalent to MacArthur & Bates Inventory focusing on word production – and a questionnaire reporting experience with multi-party situations using birth order and day-care visit as indicators since sibling children were found to learn in a different environment (Dunn & Shatz, 1989). Finally, a list of all the words of vital importance for the study was given, and the parent had to check whether the child already understood or actively used them.

3.1 Participants

A sample of 36 children aged 25 through 28 months (M = 25.8, SD = 1.2) participated in this experiment. All children were native German speakers and lived in Bielefeld and surroundings. Children received a picture book and a rubber duck for their participation.

Of the 36 children (17 girls, 19 boys) who participated, 6 (2 girls, 4 boys) had to be excluded due to fussiness (2 boys) or non-compliance (2 girls, 2 boys). The sample, therefore, consisted of 30 children, 15 boys and 15 girls. 16 were firstborns and 14 were secondborns.

3.2 Stimuli

We operationalized word learning by providing words of different word classes. The referents were different pieces of jewelry, color adjectives denominating less common colors, and number words denominating different set sizes (see Figure 2).

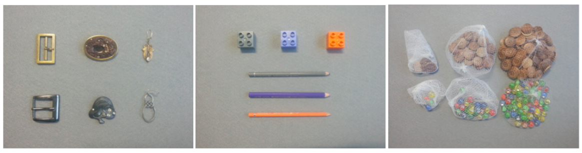


Figure 2: Stimuli for the presentation (above) and transfer (below) of nouns (left), color adjectives (middle), and number words (right).

For the acquisition of nouns, we chose labels that the children were unlikely to know, namely German words for different pieces of jewelry such as Ohrring (earring), Brosche (brooch) and

Gürtelschnalle (belt buckle). One set of these items was used to teach the words to the children, and another was used to test whether the children were able to transfer their newly acquired knowledge to another exemplar of the same object class (see Figure 2). Transfer objects differed in shape, color and size. The second word class, color adjectives were less common colors such as lila [lilac], grau [gray] and orange [orange]. During the teaching phase, colors were presented in the form of building blocks; for testing, we used crayons. We also taught children words for numbers. We chose number words such as vier [four], zwölf [twelve], and hundert [hundred] to denominate different quantities of objects. For the objects in the teaching phase, the different sets were presented using nets containing different quantities of identical wooden butons. For the transfer task, the child was presented with nets containing marbles.

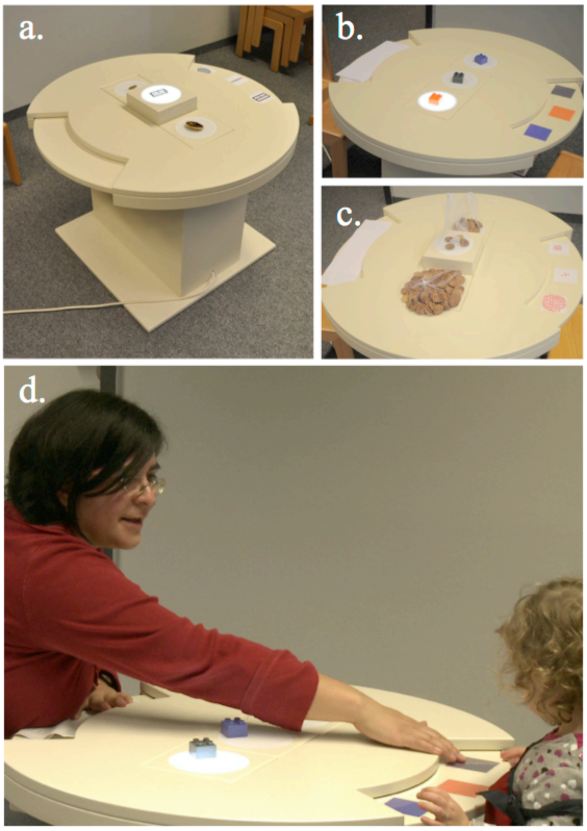


Figure 3: (a) The specifically designed table (b) the three areas in the middle of the table can be lighted up or (c) elevated in order to make the object salient; (d) on this display that the experimenter is touching, the child learned to nonverbally pick the right answer (for each word class, a different display was used as this part of the table can be rotated by the experimenter.

We randomized the ordering of the words, the ordering in which they were taught, and the position on the table on which they were presented. Each parent was asked to fill in a questionnaire during the warm-up phase that asked whether the child already knew certain words. Only when the child was reported to already know the target word, the randomization changed ad hoc.

The objects were presented on a specifically designed table (see Figure 3). The table display was used for both familiar and unfamiliar conditions as a presentation background. In the unfamiliar condition, however, the table made it possible that some elements of an interaction were unfamiliar: (a) the object was highlighted by lighting up or elevating it and (b) for the child's answer, a display was provided with featured symbols of the objects: For the noun-learning task, the display showed stylized pictures of the objects; for the adjective-learning task, the display was equipped with color patches and for the number-learning task (see Figure 3b), the pictures displayed different amounts of red dots corresponding to the numbers to be taught (see Figure 3c). These displays could be changed smoothly during the session by rotating a part of the table on the experimenter's side.

3.3 Procedure

We adopted Pepperberg's model/rival training (Pepperberg, 2002) creating a predesigned question-answer-routine. This routine contained reinforcing and corrective feedback. In both experimental conditions, the direct and the indirect teaching situation, children heard the new word five times before being testing children's learning effects that was measured using production and comprehension tests. In the tests, comprehension was defined as the child's ability to transfer the learned word to new objects. Thus, for our protocol, unlike that of Akhtar and colleagues (2001), it was not sufficient to identify the same object out of a random set of objects. Instead, children were required to use their knowledge to identify another object of the same type. As the study by Akhtar and her colleagues (2001) had shown that children – in contrast to Pepperberg's parrot – did not depend on role reversal to learn new words, we desisted from including role reversal in our experimental design, i.e. the model (the second experimenter) acted only as a learner. The whole procedure lasted 30–40 minutes with the word learning part taking ca 5 minutes.

Warm-up and pretest

After children arrived at our lab, the experimenter first engaged with the child in a simple jigsaw puzzle. Next, the experimenter tested whether the child understood the pragmatic implications of simple requests. Here the experimenter presented the child with a tray holding three objects: a train, a Playmobil® girl and a Playmobil® horse and asked the child to hand over the objects, one at a time. To make the experimental conditions comparable, we developed a script including utterances, gaze direction and gestural behavior of the experimenter(s).

Teaching

Children were taught three words from different word classes.

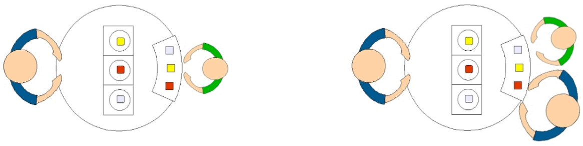


Figure 4: The two experimental conditions: unfamiliar direct versus unfamiliar indirect teaching.

As can be seen in Figure 4, in the direct teaching condition, the child was seated at a table facing experimenter 1 who acted as a tutor. In the indirect teaching condition, the child was seated at a table facing experimenter 1, who acted as a tutor, and next to experimenter 2, who acted both as a model for the child's behavior and a rival for the attention of experimenter 1. In the indirect teaching scenario, experimenter 1 reacted to the child as little as possible. In both conditions, experimenter 1 focused on his conversational partner – the child in the direct and experimenter 2 in the indirect teaching condition – and started the question-answer-routine by pointing to the object in question and asking for its name. Then, the correct name was given (either by experimenter 1 or 2 – depending on the condition) which was followed by a positive, reinforcing feedback including a reward consisting in the possibility for the learner to explore the object. Next, the routine was repeated, but this time, the answer was incorrect and was thus followed by a negative, corrective feedback. The verbal contribution by experimenter 1 in the direct teaching condition corresponded to the contribution provided by experimenter 1 and 2 in the indirect teaching condition; the child heard the new target word 5 times (3 in positive and 2 in negative formulations). After teaching, experimenter 1 proceeded to test the child's learning success.

Testing

After each teaching phase, experimenter 1 initiated the actual behavior production test: She turned to the child and call her or him by her/his given name. Then, the experimenter asked the child the same question as during the teaching phase (see Figure 1). Children sat in front of a display making it possible to provide an unfamiliar response protocol to the experimenter's questions. The child was expected to produce the learned behavior, i.e. they were expected to place their hand on the correct display in front of them (see 3.2 for more details). For scoring learning success using behavior production, children were given two points for correct and frame-appropriate production when they placed their hand on the correct display when asked for the label of the taught object; if they did not place their hand on the display but uttered the correct word, they got only one point for correct production, since they failed to produce the appropriate behavior; if the children either did not answer at all or answered incorrectly they were given no points.

In the word comprehension test, experimenter 1 cleared the table of all objects before placing an alternative set of objects in front of the child. Experimenter 1 took out a tray and asked the child to help her to place the objects on the tray. She then conducted the procedure that had previously been practiced during the warm-up phase, namely mixing the objects while saying *"mischen, mischen, mischen"* ("mix, mix, mix") and asking the child to hand over the object to which the noun referred or the object with the appropriate property by saying *"<name of the child> gibst du mir mal die Brosche?"* ("<name of the child>, would you give me the brooch?") while holding out the tray with the right hand and holding out her left hand palm up next to it, so the child knows that she waits to receive the object. For scoring, the child got two points for a correct and task-appropriate answer when she gave the experimenter the requested object or when she identified it by pointing to it. If the child handed over all objects beginning with the one the experimenter had requested, she got one point for a correct answer. This turned out to be necessary because many children seemed to have been primed by the warm-up task to hand over all items, one at a time. If the child chose not to answer at all or handed over an incorrect item or all items at once, she got no points.

4 Results

4.1 Differences between the dyadic and triadic conditions

Children in both conditions (indirect and direct teaching) on average achieved 2.5 (*SD* = 1.83) out of 6 possible points with a range from 0 to 6 in the behavior production test, and 2.1 (*SD* = 1.65) out of 6 possible points with a range from 0 to 5 in the word comprehension test.

The following Table depicts the distribution of the achieved scores:

	production			**comprehension**		
scores	**0**	**1**	**2**	**0**	**1**	**2**
direct teaching	26	9	10	26	3	16
indirect teaching	22	3	20	29	4	12

Table 1: Children's performance according to the score distribution; each child participated in 3 trials.

A nonparametric Mann-Whitney test showed no significant differences between boys (*N* = 15) and girls (*N* = 15) in their overall performance either in behavior production (*U* = 100.5, *p* = 0.62) or in word comprehension (*U* = 95.5, *p* = 0.48). The children in the two experimental groups direct teaching (*N* = 15) and indirect teaching (*N* = 15) did not differ in lexical development (*U* = 86.5, *p* = 0.28).

Given that the data were not normally distributed (Kolmogorov-Smirnov *df* = 15, *p* < 0.05 for word production and comprehension), nonparametric Mann-Whitney-tests were performed. Additionally, Spearman's Rank Order Correlations were computed between the children's performance and lexical development, shyness, and experience with multi-party situations operationalized by birth order and daycare experience.

Our data (see Figure 5) from the production tests revealed that when taught directly, children scored poorer (33.3 % of the possible correct responses) than children taught indirectly (50 % of the possible correct responses). In the word comprehension test, the result was reversed with children in the direct teaching condition achieving 38.9 % of correct answers and children in the indirect teaching condition scoring 31.1 % of the possible correct responses (see Figure 5).

Nonparametric Mann-Whitney-tests for overall production and comprehension showed no significant differences between children's performance in both conditions (production: *U* = 84, *p* = 0.11; comprehension: *U* = 93, *p* = 0.20, one-sided).

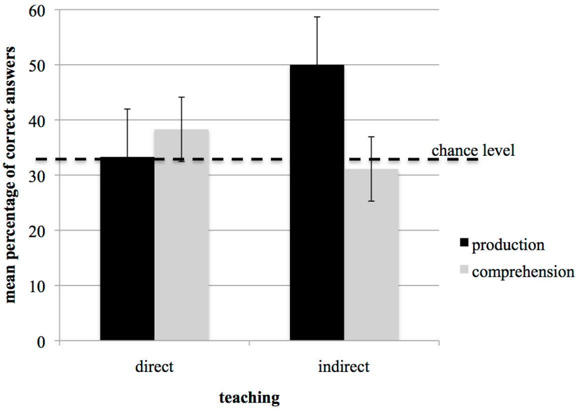


Figure 5: Overall performance in the unfamiliar direct (dyad) and indirect teaching (triad) conditions.

4.2 Learning effects

To assess the learning effects, we performed a Wilcoxon test comparing children's performance to a chance level of 33 %. We found that children in the direct condition did not differ significantly neither for production (*Z* = -0.274, *p* = 0.39, one-sided) nor comprehension (*Z* = 0.847, *p* = 0.19, one-sided). The same was true for children's comprehension in the indirect teaching condition (*Z* = -0.927, *p* = 0.463, one-sided). However, for the production, we found that children's performance in the indirect teaching scenario was significantly different than at the chance level (*Z* = -1.621, *p* = 0.052, one-sided) suggesting that only in the indirect teaching scenario, children improved their production.

In sum, the children accepted the unfamiliar frame conditions readily. They understood that they were expected to produce a response – which was elicited by addressing them with a direct question in the production test – but they also learned that uttering a word would not be the appropriate way to behave in this interaction. After all, only in 2.2 % of all cases did the children try to answer the question by producing a word rather than this novel nonverbal behavior. Thus, in the most testing trials, the children either refused to answer or applied the new behavior in trying to respond to the experimenter. In the aftermath of the experiment, they even tended to create their own non-verbal frames by placing a hand on one of the pictures placed in front of them and looking to the experimenter prompting her to utter a label and then replacing the hand on another picture and again gazing at the experimenter etc. In none of the cases, in which the children initiated these games, did they try to include speech.

4.3 Relations to language skills and birth order

Correlations of children's overall performance reported to their lexical development displayed no relation with children's performance (production: $r_s = 0.31$, $N = 30$, $p = 0.10$; comprehension $r_s = 0.09$, $N = 30$, $p = 0.63$) implying that children reported to have a more advanced lexicon performed similarly to children reported to have a less advanced lexicon.

Next, we compared the performance of children who had older siblings or visited daycare, to firstborns or children who stayed at home with their mothers, because the former are supposed to have more experience in multi-party interaction than the latter. Our correlational analyses between overall performance and birth order did not reveal any relationship of experience in multi-party interactions with task performance (production: $r_s = 0.08$, $N = 30$, $p = 0.70$; comprehension: $r_s = 0.04$, $N = 30$, $p = 0.84$). Furthermore, no significant correlations could be found for overall performance and daycare visit (production: $r_s = 0.02$, $N = 30$, $p = 0.93$; comprehension: $r_s = 0.20$, $N = 30$, $p = 0.29$).

5 Discussion

The children in this experiment presented above learned words from various word classes within an unfamiliar frame. This means that they experienced an interaction protocol with novel aspects as a new way of singling out referents and a new way of responding (placing one's hand on a display). These novel aspects of the interaction protocol differed from interactional knowledge that children had at their disposal, because commonly, a question is answered by a verbal behavior (Anselmi et al., 1986).

We expected children in the indirect teaching condition to follow the new interaction protocol and to learn the reference better because the multi-party situation presented them with a model to imitate, thereby making the expectation of how they should behave more transparent. Our results confirm our hypothesis. In general, children are able to learn new frames, i.e. a new interaction protocol from an ongoing interaction as both groups were able to apply the displayed symbols for an object or one of its characteristics (e.g., its color or amount). This achievement strongly supports the idea that when children learn words, they master many tasks concurrently (Clark, 1974). However, while one challenge consists of learning an appropriate behavior, the other – perhaps greater – challenge is to learn a new word and its concept. In our study, children not only had to acquire a concept of the new word and to bring this knowledge into the comprehension task, in which they had to pick the right example of this referent, they also had to apply the new concept within an appropriate, newly acquired nonverbal behavior in the production test. While we found no differences between the direct and indirect teaching conditions when compared the groups directly, only children in the indirect teaching scenario performed at a significantly better level than chance in their production test.

These findings put us in the position to think that when children can bring little previously acquired pragmatic knowledge to comparable teaching situations, but must acquire the pragmatics during the ongoing situation and learn a semantic content, they will perform better when exposed to the indirect teaching than those children taught directly in the production test. For the comprehension task, in contrast, the achievement in both groups was comparable.

Further, we assumed that birth order and daycare visit as operationalizations of the children's experience with multi-party interactions would enhance the advantage of the indirect over the direct teaching condition. This hypothesis could not be confirmed: The extent of experience with multi-party interactions did not influence children's performance in experiment 2 implying that all children can benefit equally from indirect teaching independently from how much experience with this kind of situations they had acquired previously. In the case of lexical development, our correlational analyses showed no significant relation to word comprehension and production tests.

Our results thus suggest that two-year old children benefit most from modeling taking place in indirect teaching conditions when the pragmatic frame is unfamiliar and thus the learning task puts high cognitive demands on the child. In such cases, children's ability to draw on already acquired interactional behavior is limited, and they seem to make use of an imitation mechanism that allows them to (a) to pick up the pragmatic information provided in the teaching situation and (b) keep the interaction going by simply copying the interactional behavior previously displayed by the model and thus accomplish the task. The reason why we think that a cognitively less demanding mechanism of imitation is applied here are our obtained results in the comprehension test: Although in an unfamiliar frame, these indirectly taught children showed a better productive behavior, they did not perform better in a comprehension task suggesting that their concept of the presented new word remains weak and linked to a specific action. We suggest that taking advantage of indirect teaching does not mean that children achieve a deeper understanding of the object-label match but it allows them to stay further engaged in the ongoing situation (see behavior of impaired adults in Wrede et al., 2010), thereby prolonging the chance to learn from it. This is in line with adult research suggesting that overhearing does not lead to a better understanding. On the contrary, when addressed directly, adult participants demonstrated a much more accurate understanding of an instruction (Schober & Clark, 1989). We can extend the findings with our data from children suggesting that imitation does not seem to substitute for or boost cognitive processes.

As to the question of whether the acquisition of pragmatic frames is a prerequisite or an integral part of word learning, we see in our data that while the knowledge of frames is crucial for word production, it does not enhance the word comprehension. Thus, the acquisition of pragmatic frames seems to be an integral part of the learning process, needed for the emergence of a solid word concept.

Acknowledgments

The research reported in this paper took place within Juana Salas Poblete's dissertation project that was supported by Honda Research Institute Europe. We are very grateful to Irene Pepperberg and the reviewers for their helpful comments on an earlier draft of this paper.

References

Akhtar, N. 2005. The robustness of learning through overhearing. *Developmental Science,* 8: 199–209.

Akhtar, N., & Herold, K. 2009. Pragmatic development. In J. B. Benson & M. M. Haith (Eds.), *Language, Memory, and Cognition in Infancy and Early Childhood* (pp. 399–409). Oxford: Elsevier.

Akhtar, N., Jipson, J., & Callanan, M. A. 2001. Learning words through overhearing. *Child Development*, 72: 416–430.

Anselmi, D., Tomasello, M. & Acunzo, M. 1986. Young children's responses to neutral and specific contingent queries. *Journal of Child Language*, 13: 135–144.

Bandura, A. 1971. *Social learning theory.* New York: General Learning Press.

Barton, M. E., & Tomasello, M. 1991. Joint attention and conversation in mother-infant-sibling triads. *Child Development,* 62: 517–529.

Booth, A.E., McGregor, K.K., & Rohlfing, K. J. 2008. Socio-pragmatics and attention: Contributions to gesturally guided word learning in toddlers. *Journal of Language Learning and Development*, 4: 179–202.

Bruner, J. 1983. *Child's Talk.* Oxford: Oxford University Press.

Carpenter, M., Tomasello, M., & Striano, T. 2005. Role reversal and language in typically developing infants and children with autism. *Infancy*, 8: 253–278

Clark, R. 1974. Performing without competence. *Journal of Child Language*, 1: 1–10.

Crozier, R. W., & Hostettler, K. 2003. The influence of shyness on children's test performance. *British Journal of Educationla Psychology*, 73: 317–328.

Dunn, J., & Shatz, M. 1989. Becoming a Conversationalist Despite (or Because of) Having an Older Sibling. *Child Development*, 60: 399–410.

Floor, P., & Akhtar, N. 2006. Can 18-Month-Old Infants Learn Words by Listening In on Conversations? *Infancy*, 9: 327–339.

Fogel, A. 1993. Two principles of communication: Co-regulation and framing. In J. Nadel & L. Camaioni (Eds.), *New perspectives in early communicative development.* London: Routledge.

Franco, F., & Butterworth, G. 1996. Pointing and social awareness: declaring and requesting in the second year. *Journal of Child Language*, 23: 307–336.

Gampe, A., Liebal, K. & Tomasello, M. 2012. Eighteen-month-olds learn novel words through overhearing. *First Language*, 32: 385–397

Grimm, H., & Doil, H. 2006. *Elternfragebögen für die Früherkennung von Risikokindern (ELFRA-2).* Göttingen: Hogrefe.

Heath, S. B. 1983. *Ways with words.* Cambridge: Cambridge University Press.

Herold, K. H., & Akhtar, N. 2008. Imitative learning from a third-party interaction: Relations with self-recognition and perspective taking. *Journal of Experimental Child psychology*, 101: 114–123.

Ninio, A., & Snow, C. E. 1996. *Pragmatic Development.* Boulder, Colo: Westview Press.

Ochs, E. 1986. Variation and error: A sociolinguistic approach to language acquisition. In D. Slobin (Ed.), *The cross-linguistic study of language acquisition* (Vols. 1-2, Vol. 1, pp. 783–837). Hillsdale, N.J.: Erlbaum.

Oshima-Takane, Y. 1988. Children learn from speech not addressed to them: the case of personal pronouns. *Journal of Child Language*, 15: 95–108.

Oshima-Takane, Y., Goodz, E., & Derevensky, J. L. 1996. Birth Order Effects on Early Language Development: Do Secondborn Children Learn from Overheard Speech? *Child Development*, 67: 621–634.

Pepperberg, I. M. 1997. Social influences on the acquisition of human-based codes in parrots and nonhuman primates. In C. T. Snowdon & M. Hausberger (Eds.), *Social influences on vocal development* (pp. 157–177). Cambridge UK: Cambridge University Press.

Pepperberg, I. M. 2002. *The Alex studies: cognitive and communicative abilities of grey parrots.* Cambridge, Mass.: Harvard University Press.

Pepperberg, I. M., & Sherman, D. V. 2002. Use of two-trainer interactive modeling as a potential means to engender social behavior in children with various disabilities. *International Journal of Comparative Psychology*, 15: 138–153.

Pruden, S. M., Hirsh-Pasek, K., Michnick Golinkoff, R., & Hennon, E. A. 2006. The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development*, 77: 266–280.

Pye, C. 1986. Quiché Mayan speech to children. *Journal of Child Language*, 13: 85–100.

Schieffelin, B. B. 1986. The acquisition of Kaluli. In D. Slobin (Ed.), *The cross-linguistic study of language acquisition* (Vols. 1-2, Vol. 1, pp. 525–593). Hillsdale, N.J.: Erlbaum.

Schober, M. F., & Clark, H. H. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21: 211–232.

Scollon, R., & Scollon, S. B. K. 1981. *Narrative, Literacy and Face in Interethnic Communication. Advances in Discourse Processes.* Norwood, N.J.: Ablex Publishing Corporation.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition.* Cambridge, Mass., London: Harvard University Press.

Tomasello, M. 2003. Constructing a language. Cambridge, Mass.: Harvard University Press.

Tomasello, M., & Akhtar, N. 2000. Five questions for any theory of word learning. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 179–186). Oxford: Oxford University Press.

Wrede, B., Kopp, S., Rohlfing, K., Lohse, M., & Muhl, C. 2010. Appropriate feedback in asymmetric interactions. *Journal of Pragmatics*, 42(9): 2369–2384.

The Roadsigns of Communication

Tania Rojas-Esponda
Stanford University

Abstract

Special intonation and discourse particles can act as pragmatic roadsigns that signal specific moves in conversation. By making the nature of a conversational move explicit, these devices can aid in pragmatic processing. I make this idea precise using a Question under Discussion framework. Several case studies are presented.

1 Introduction

Questions under Discussion (Ginzburg, 1995a; Ginzburg, 1995b; Roberts, 1996) have proved to be a powerful concept for capturing the structure of conversation. Among other things, QUDs have demonstrated their usefulness for the understanding of focus (Roberts, 1998; Geurts and van der Sandt, 2004; Büring, 2003), anaphora resolution (Roberts, 2003; Clark and Parikh, 2007; Schoubye, 2009), speech acts (Roberts, 2004), scope resolution (Zondervan et al., 2008), presupposition (Thomason et al., 2006), and quantifier domain restriction (Malamud, 2006).

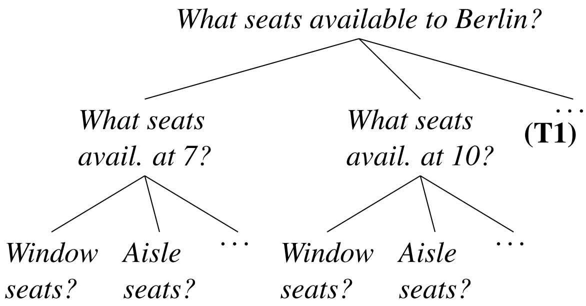
What the QUDs are in a given conversation is a matter that will be inherently interactional, and which is subject to negotiation between speakers. In this paper I discuss explicit devices that speakers can use to signal their views and preferences about the QUDs. These include intonation (Roberts, 1998; Büring, 1999; Büring, 2003) and discourse particles (Beaver and Clark, 2008; McCready, 2006; Davis, 2009; Eckardt, 2007). By making specific conversational moves overt, these devices act as 'pragmatic roadsigns.' This is especially useful when a conversational move is unexpected or could be construed as uncooperative; in those cases, these expressions or intonational devices can help speakers more effectively align their mental maps of the conversation. This paper presents a novel view of how QUD hierarchies are structured and then discusses three such roadsigns in more detail.

2 Discourse as QUDs

Imagine a scenario in which a traveler wants to get a flight to Berlin. If the traveler asks *Are there window seat tickets to Berlin at 7:00?*, then a simple *No* answer is not as helpful as one of the following replies:

1. *There are no seats (at all) for the 7:00 am flight to Berlin.* [no seats as opposed to just no window seats]
2. *The next available seats are for 10:00.*
3. *There are no seats/tickets to Berlin today at all.*

The answers in 1-3 above are helpful assuming that the most important thing for the traveler is to get some seat to Berlin today. A possible goal or question hierarchy is shown below. There, the overarching question is *What seats are available to Berlin?*, with subquestions about the seats at different times of day. These can furthermore have subquestions like *Are there window seats to Berlin available at 7:00?...*



(T1)

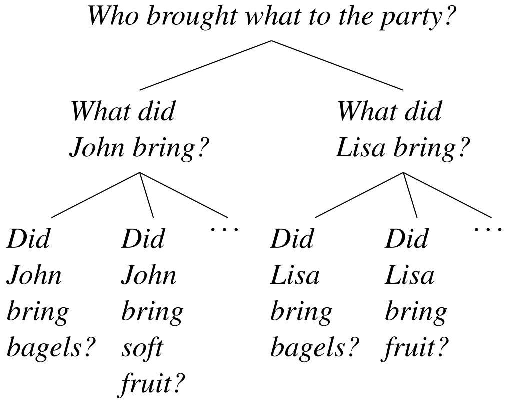
If the travel agent infers that the question asked by the traveler was serving the bigger question *What are the seats available to Berlin at 7:00?*, then instead of just passing on the information that there are no window seats on that plane, she can give an answer that gives more information towards solving the question of what seats (window or otherwise) are available on that plane and utter 1. Or, similarly, if the seats at 7:00 are all gone, she could offer information about flights to Berlin at other times, which also helps towards the answer of the bigger question of what seats are available to Berlin. Finally, the agent can utter 3, thereby resolving the highest question in one go.

By representing the problem structure that the traveler is facing in terms of a question hierarchy, we can make sense of different conversational strategies. Answers 1-3 are all more effective in advancing the traveler towards solving their bigger goal of getting a flight to Berlin than the direct answer *No*. Even though answers 1-3 are very effective given this problem structure, they do depart from the most direct answer. If a polar question *p or not p?* signals a request for either one of the answers *p* or *not p*, then 1-3 are marked with respect to this most expected type of answer. Such marked or less expected moves are often signalled using special intonation or discourse particles. This idea will be illustrated in section 4.

3 The S-tree formalism

3.1 S-trees

Roberts (1996), Groenendijk (1999) and Büring (2003) developed the idea of a hierarchy of questions. When going about a difficult problem, interlocutors may divide it into pieces that are simpler and attempt to solve these instead. In terms of questions, this means that each question is divided into a collection of related questions which, when answered, provide just enough information to answer the original question. An example of how this subdivision into simpler questions might look is shown below:



(T2)

The S-tree formalism differs from the question hierarchies in Roberts (1996) and Groenendijk (1999) in that it allows more flexibility with respect to presuppositions. Namely, in S-trees presuppositions can be added or removed at different levels of the tree. Why this is desirable can be seen in tree (T2). The superquestion *Who brought what to the party?* might have the presupposition that each person (in a contextually relevant set) brought something to the party. Its daughter questions, for instance *What did John bring?*, do not have this presupposition, but possibly a weaker presupposition that John brought something. As we go one level lower in the tree, we get questions like *Did John bring pizza?* that have no presupposition.

Even though presuppositions present in a parent question can be missing in the child questions or vice versa, the presence or absence of presuppositions can have important effects on the conversation, as well as on the use of discourse particles (see 4.3). To allow us to track presuppositions, we relax the definition of questions as partitions of the entire world set (Groenendijk and Stokhof, 1982; Groenendijk and Stokhof, 1984; Lewis, 1988). Let W be the set of possible worlds. A question is represented as a partition of some proper or improper subset of W. Equivalently, a question is defined as a symmetric, transitive binary relation Q on W. When Q is reflexive, the question is a partition on the whole world set. When Q is not reflexive, we can think of it as a partition on a proper subset $S \subsetneq W$. In this case S corresponds to the presupposition of the question Q.

An *answer* to a question Q will be defined as in Groenendijk (1999), as an assertion that picks out an integer number (zero or more) of full cells of Q. A *complete* or *full answer* is a question that picks out a single cell of Q. A *partial answer* is an answer that picks out two or more full cells of Q.

The goal is to define S-trees in such a way that they are question trees with the following properties:

1. If we answer all the daughter questions of a parent question Q, we arrive at the answer to Q (or the statement that Q is not answerable).
2. If we provide a full answer to the parent question Q, we get an answer to each daughter question (or the statement that a daughter question is not answerable).

3. A question in an S-tree can introduce a presupposition not present in its parent question.
4. A daughter question may lack a presupposition present in the parent question.

To characterize the relationship between children nodes and mother nodes in S-trees, I define the concepts of *narrowing* and *combination*. Roughly speaking, a question q *narrows* Q if q only raises issues raised by Q. Intuitively, two questions q_1 and q_2 *combine* to give Q if answering q_1 and q_2 is a way to get the answer to Q. Using this terminology, S-trees are question trees that satisfy the following:

- Each child node *narrows* its parent node
- The children of a node Q (intersected with the presupposition of Q) *combine* to produce Q

First we define the *completion* of a question: a question is *completed* by adding a cell containing all worlds not already in the question. When applied to a statement p it results in a polar question: p or not p?

Definition 1 (completion)**.** The *completion* $\bar{q}$ of q is

$$\bar{q} \equiv q \cup \{\, \langle v, w\rangle \mid \langle v, v\rangle \notin q \;\& \langle w, w\rangle \notin q \,\}.$$

The narrowing relation generalizes the notion of subquestion from Groenendijk (1999).

Definition 2 (narrowing)**.** If q and Q are questions in $W \times W$, then we say that q *narrows* Q ($q \prec Q$) if and only if, for every pair of worlds $(v, w) \in W \times W$,

$$\langle w, w\rangle \in q \;\text{ and }\; \langle v, w\rangle \in Q \implies \langle v, w\rangle \in q$$

and

$$\langle v, v\rangle \notin Q \;\text{ and }\; \langle w, w\rangle \notin Q \implies \langle v, w\rangle \in \bar{q}$$

If q narrows Q then q only raises issues raised by Q. Any answer for q partially answers Q, or shows Q is invalid. Any answer to Q completely resolves q.

The definition of a *combination* follows:

Definition 3 (combination)**.** The *combination* $q_1 \square q_2$ of q_1 and q_2 is $q_1 \square q_2 \equiv (\bar{q_1} \cap \bar{q_2}) \cap (q_1 \cup q_2)$.

The combination of two questions q_1 and q_2 is the most general question that can be answered by resolving q_1 and q_2.

The *smash* of a question q is the statement that q can be answered. Intuitively, it is the presupposition implicit in q.

Definition 4 (smash)**.** The *smash* $\hat{q}$ of q is defined by

$$\hat{q} \equiv \{\, \langle v, w\rangle \in W \times W \mid \langle v, v\rangle \in q \;\& \langle w, w\rangle \in q \,\}$$

We are now ready to define S-trees.

Definition 5 (S-tree)**.** A strategy tree, or S-tree, is a question tree satisfying the following:

Every child q of a node Q satisfies the relation $q \prec Q$, and the children $q_1, \ldots, q_k$ of Q satisfy $Q = \square_{i=1}^{k} q_i \cap \hat{Q}$.

S-trees are useful tools to understanding the flow of conversation. In order to capture the evolution of a conversation, we must establish rules for how one may move in an S-tree. The rules of traversal below build on Roberts (1996):

Definition 6 (Rules of traversal)**.** We may proceed from a node to a sister node or to a child node. However, we may only move to a parent or ancestor node if we do one of the following:

(i) Resolution: Resolve parent node Q by providing a full answer to it

(ii) Doubting: Show the parent node Q to be unanswerable by stating the negation of its presupposition, $\neg\hat{Q}$.

Alternatively, we can try to move upwards from a question Q by forming the polar questions corresponding to the *resolving* and *doubting* moves. A valid move upwards in a tree is called an *ascending* move.

The S-tree formalism presented here is based on Rojas-Esponda (To appear a).

3.2 Comparison with other QUD theories

Two ways in which the S-tree formalism differs from the question hierarchies in Roberts (1996) and Groenendijk (1999) are the explicit tracking of presuppositions and the freedom to add and remove presuppositions at different levels of the tree.

Keeping track of presuppositions is achieved by letting a question be a partition on a subset of the world set. A question in the S-tree formalism is a symmetric, transitive binary relation R on W instead of an equivalence relation as in Groenendijk and Stokhof (1984) and Groenendijk (1999). By allowing some world pairs $\langle w, w \rangle$ to be excluded from the relation R (i.e. by not requiring reflexivity) one can model that some questions have non-trivial presuppositions and thus are only answerable on a proper subset $S \subsetneq W$. This does not mean that a question defined on a proper subset $S \subsetneq W$ asserts S. Rather, the truth of S can be negotiated among speakers. The greater point here is that allowing questions to partition subsets allows one to make sense of the notion that not only assertions, but also questions can be challenged (see (Rojas-Esponda, To appear a) and (Rojas-Esponda, To appear b) for why this is important for the particles *überhaupt* and *doch*, respectively).

Another difference from, e.g. the subquestion relation of Groenendijk (1999), is that child nodes are only required to narrow the parent node and combine to give the parent node. This allows the flexibility of adding or removing presuppositions as you move down one level in the tree, from a parent to a child node. Why this is desirable is shown below:

Another theory of QUD-trees, called D-trees, was developed by Büring (2003). I explain below why D-trees don't share one key feature with the formalism of S-trees or the formalisms of Roberts (1996) and Groenendijk (1999), namely that of being information-theoretically hierarchical.

Büring uses two types of restrictions in defining the class of D-trees. The restrictions that are based just on information-theoretic content are shown below:

Definition 7. A is an answer to Q if A shifts the probabilistic weights among the propositions denoted by Q.

Definition 8. q is a daughter question of Q iff at least one answer to q is an answer to Q.

Unraveling these definitions, we get that q is a daughter question of Q iff there exists at least one proposition a_1 that shifts the probabilistic weights of both q and Q. But this restriction is symmetric in q and Q. Therefore, without the other constraints used by Büring (based on CT- or F-marking), we would get that q is a daughter question of Q if and only if Q is a daughter question of q. This would not give us a directed structure, but due to its symmetry would yield something more akin to a cluster or graph.

4 Case Studies

In this section I will present a number of case studies that illustrate how we can use special language resources, like intonation or discourse particles, in order to obtain insights into how interlocutors are negotiating issues in discourse.

4.1 Intonation

Intonation is an important pragmatic roadsign, as different choices of intonation can give crucial clues as to what QUDs interlocutors are entertaining. Moreover, intonation can either mark congruence with a question asked or signal a change to a different question under discussion. These ideas were developed in Roberts (1998), Büring (1999) and Büring (2003), among other places.

Consider the following conversation:

<u>Conversation 1</u>

(i) *A: Who brought bagels?*

(ii) *B: SONJA brought bagels.*

In conversation (C1), B's answer has focus marking that is congruent with the question asked. Since the proper name Sonja is stressed, accounts of focus as generating alternatives (Jackendoff, 1972; von Stechow, 1981; Rooth, 1985; Taglicht, 1984), yield the set of propositions *X brought bagels.*[1] Importantly, B's focus marking seems to be coherent with the question raised by A.

Compare this to the following exchange:

<u>Conversation 2</u>

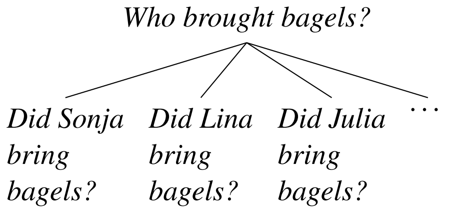
(i) *A: Did Sonja bring bagels?*

(ii) *B: LINA brought bagels.*

The declarative *LINA brought bagels,* with contrastive topic accent on LINA, is not an answer to to A's question about Sonja. What is the rationale behind B's reply? Instead of merely answering the polar question asked, B offers information about Lina. This makes sense if B sees A's question not as the only QUD in the conversation, but sees A's question as serving another, larger QUD. This could be the question *Who brought bagels?*. Thus a tree incorporating both what A and what B said could be the following:

---

[1] It is important that B's accent is a focus accent, not a topic accent. This was shown by Büring (1999).



(T3)

Looking at tree (T3), we can see why B's answer in (C2) is strategic. However, B is not providing the exact information requested by A's polar question about whether Sonja brought bagels. The contrastive topic accent used by B is a clue that makes overt B's less expected conversational move (in this case, a move to answer a sibling question tied to A's question by a common superquestion).

4.2 The discourse particle *noch*

In this section I discuss a further roadsign, namely the German particle *noch* as analyzed by Eckardt (2007).[2]

The following is an example of a discourse use of *noch*, from Eckardt (2007).

Conversation 3

Tick kann schwimmen, und TRICK kann noch *schwimmen, (aber) Track kann nicht schwimmen.*

Tick can swim, TRICK can *noch* swim, (but) Track cannot swim.

Here, the question under discussion seems to be *Who can swim?* or *Which of Donald Duck's nephews can swim?* This question could be thought of as having three subquestions (in the sense of Groenendijk (1999)), *namely Can Tick swim?*, *Can Trick swim?* and *Can Track swim?* The first is answered in the positive, the second is answered in the positive as well, and the third is answered in the negative. This forms the basis for Eckardt's analysis. Namely, Eckardt proposes that *noch* in assertions can be used when we have a series of assertions that are answers to subquestions of a larger QUD and when all of the preceding assertions in this series have been 'yes'-answers to their corresponding questions.

In her own words:

> *noch* in assertions can occur in the n-th assertion of an ongoing strategy iff $n > 1$ and if all previous assertions pertained to the current question under debate positively (i.e. were a 'yes' answer to the local subquestion).

In order to account for *noch* in questions, Eckardt uses the notion of a remnant question (see also Büring (2003)). Roughly speaking, a remnant question is obtained when a question Q such as *Who can swim?* is addressed via a partial resolution, e.g. *Pat can swim*, and a question *Who else can swim?* that asks for the part of the question that remains unaddressed. This notion is handy as *noch* can be used in exactly this kind of question:

Conversation 4

Lucy kann schwimmen. Wer kann NOCH schwimmen?

Lucy can swim. Who else can swim?

Eckardt then analyses *noch* in questions as follows:

> Use of *noch* in questions: A question q licenses *noch* iff (a) it is a remnant question and (b) it is dominated by a question Q such that there are assertions between Q and q, and all assertions between Q and q are positive answers to Q.

I will sketch how Eckardt's account of *noch* can be captured within an S-tree formalism. The formalism in Eckardt's account is called a Question Answer Discourse — QAD. Let's start with the answerhood definition in the QAD formalism. In the S-tree formalism an (informative) answer to a question Q is an assertion that eliminates one or more full cells from Q. In QAD, the answerhood relation is more permissive: it allows overanswering. In terms of partitions, this would mean that an answer could pick out parts of cells, not just entire cells. However, Eckardt does not make use of this freedom in her *noch* examples. Thus, one could prohibit overanswering and the examples would stay intact. In fact, overanswering is not desirable for *noch*. Let's say we have a question Q = *Who can swim?* with subquestions *Can Tick swim?*, *Can Trick swim?* and *Can Track swim?* Let's also assume we allow overanswers. Because *Trick is a world champion swimmer* entails *Trick can swim*, we should be able to say:

---

[2] Focus marking is relevant for the particle *noch*, but I gloss over the distinctions between focused and unfocused *noch* here for reasons of brevity. For a fuller description, see Eckardt (2007).

Conversation 5

# Tick kann schwimmen, TRICK ist noch ein Weltmeister im Schwimmen, aber Trick kann nicht schwimmen.

# Tick can swim, TRICK is *noch* a world champion in swimming, (but) Track cannot swim.

However, (C5) is not felicitous. Thus, it is not desirable to allow overanswering in a *noch* formalism. However, once overanswering is eliminated and we allow only full cells to be removed, then the notion of subquestion from Eckardt (2007) becomes essentially a special case of the narrowing relation from 3.1.

Now, if instead of requiring q to be a subquestion of Q, one merely requires it to narrow Q, then one gains freedom with respect to introduction or elimination of presuppositions. I claim this does not hurt the *noch* analysis, and is even desirable:

Let Q be the question *Who can swim?* We may want to cut this into the three subquestions *Can Anna swim?, Can the neighbor's daughter swim?* and *Can Lisa swim?* The second question has a presupposition about the existence of a neighbor and a daughter of this neighbor. The particle *noch* can be used here:

Conversation 6

Wer kann schwimmen?

Who can swim?

Anna kann schwimmen, die Tochter des Nachbarn kann noch schwimmen und Lisa kann noch schwimmen.

Anna can swim, the daughter of the neighbor can noch swim and Lisa can noch swim.

By the arguments above, S-trees preserve the essential features of the QAD framework while capturing some of the data better. S-trees also handle assertions (a special type of question with just one cell) naturally. In the QAD framework, there are two different ways that a parent question can split into child nodes. One involves a splitting into two questions and the other splitting into an assertion plus a remnant question. These have to be defined separately because the splitting into questions is defined via answerhood and answerhood is not defined for an assertion. In the S-tree framework, the relation between parent and child nodes is defined using narrowing and combination, neither of which directly use the notion of answerhood. Thus, no extra work is needed to incorporate assertions into S-trees. In an S-tree, we can define a remnant question simply as the sole right sibling to an assertion node.

The particle *noch* provides hearers with information about the QUD structure of a conversation. Namely, *noch* signals that the utterance containing it is a positive answer in a sequence of (positively answered) sibling questions tied together by a common superquestion. The particle is especially useful when the hearer might have thought that all positive answers to the subquestions had already been listed. For instance, in (C6), a listener might have thought that the question of who can swim was exhaustively answered by *Anna can swim.* , an expectation which is overwritten by the two *noch*-clauses. By making overt the QUD move that is being made, the speaker can facilitate comprehension for the listener. Compare this to a similarly structured dialogue without *noch*.

Conversation 7

Wer kann schwimmen?

Who can swim?

Anna kann schwimmen. Lisa kann schwimmen.

Anna can swim. Lisa can swim.

In conversation (C7) there is a greater risk that the information about Lisa will not be understood as an additional piece of information, but instead as a correction to the assertion that Anna can swim.

4.3 The discourse particle *überhaupt*

I argue that the German particle *überhaupt* acts as a conversational roadsign, namely by signalling a move to a higher question under discussion in a hierarchical QUD strategy. The discussion will show that in order to deal with *überhaupt*, we need to rely on the mechanisms for handling presuppositions discussed in section 3. For instance, a looser notion than that of subquestion is needed, namely *narrowing* (see 3.1). I argued in 4.2 that the looser notion of narrowing is also useful for analyzing the particle *noch*.

The particle *überhaupt* has several, apparently disparate uses (König, 1983; Anderssen, 2006), and focus plays a role. Here I summarize a unified account that considers *überhaupt* as signaling a move to a higher Question under Discussion. For more details, see Rojas-Esponda (To appear a).

The uses of *überhaupt* are outlined below.

Focused *überhaupt* is used in a statement which generalizes previous statements in the dialogue:

Conversation 8

(i) *A: Verkaufen Sie Marmorkuchen?*
A: Do you sell marble cake?

(ii) *B: Nein.*
B: No.

(iii) *A: Verkaufen Sie Schokoladenkuchen?*
A: Do you sell chocolate cake?

(iv) *B: Wir verkaufen* ÜBERHAUPT *keinen Kuchen.*
B: We sell ÜBERHAUPT no cake.

The last utterance can be paraphrased as *We don't sell any cake at all.* Once this is uttered, the line of interrogation about what cake interlocutor B sells is terminated because the answer to every question (*No*) is implied by statement (C8.iv). Alternatively, if B had merely said he sells no chocolate cake, A could have replied *Verkaufen Sie* ÜBERHAUPT *Kuchen?* (Do you sell ÜBERHAUPT cake?), which can be paraphrased as *Do you sell any cake at all?*.

Below is an example illustrating the use of unfocused *überhaupt*.

Conversation 9

(i) *A: Möchtest du ein Glas Wein?*
A: Would you like a glass of wine?

(ii) *B: Nein, Danke.*
B: No, thank you.

(iii) *A: Hättest du gerne ein Bier?*
A: Would a beer appeal to you?

(iv) *B: Nein. Ich trinke überhaupt keinen Alkohol.*
B: No. I drink *überhaupt* no alcohol.

The last sentence can be paraphrased by *I actually don't drink alcohol.* As in (C8), *überhaupt* here has the effect of terminating a line of inquiry by generalizing over it. But in its unfocused form, *überhaupt* plays an additional role, namely that of invalidating a presupposition. A question equivalent of this usage also exists. If B had merely said he wants no beer, A could have replied *Trinken Sie* überhaupt *Alkohol?* (Do you even drink alcohol?).

Finally, *überhaupt* may be used with a universal quantifier or scalar predicate. In this use, *überhaupt* is always focused.

Conversation 10

(i) *A: Wie war das Wetter, als du in Rom warst?*
A: How was the weather when you were in Rome?

(ii) *B: Das Wetter war gut.*
B: The weather was good.

(iii) *A: Wie waren die Leute?*
A: How were the people?

(iv) *B: Die Leute waren sehr nett. Es war* ÜBERHAUPT *(alles) sehr schön in Rom.*
B: The people were very nice. It was ÜBERHAUPT very nice in Rome.

The last sentence can be paraphrased as *It was overall very nice in Rome.* This use also has a corresponding question form: *"War* ÜBERHAUPT *(alles) schön in Rom?"* (Was generally everything nice in Rome?).

When confronted with a series of questions that appear to be subquestions of a larger question Q, interlocutors can use *überhaupt* to move to the higher question Q or even to a superquestion of Q.

We use *überhaupt* if we doubt a higher question makes sense, or to resolve it directly rather than by answering subquestions.

In the conversation about drinks (C9), the use of unfocused *überhaupt* signals that a presupposition of the superquestion might have been invalid, suggesting that the superquestion was something like *What is the alcoholic drink that you want?* In the Rome conversation (C10), on the other hand, B may understand that A is asking a series of subquestions of the larger question *What were things like in Rome?* and B decides to answer this higher question directly, in an utterance that includes *überhaupt*.

The meaning of überhaupt: After utterances U, interlocutor i may felicitously utter *überhaupt* (q) only if q is an ascending move in $S(i,U) \in \mathrm{Stra}(U)$. Thus the presence of *überhaupt* in q signals that q is an ascending move in $S(i,U)$.

Notation used in the denotation above: q stands for either a question or a declarative sentence. Say the set of utterances so far in the conversation is U. Let $\mathrm{Stra}(U)$ be the set of all compatible strategy trees. The tree 'favored' by each interlocutor i among the set of trees $\mathrm{Stra}(U)$ is denoted $S(i,U)$. Intuitively, the tree $S(i,U)$ is interlocutor i's view of how the discourse is organized.

By using *überhaupt* a speaker can make explicit a conversational move that deviates from the exact information requested. For instance, in (C8), the explicit questions concerned just marble cake and chocolate cake, so the answer that B provides deviates from an answer that gives strictly the information requested and nothing else. B's answer violates the notion of relevance as defined in (Groenendijk, 1999). Yet B's answer is strategic, as B is trying to help A answer what she presumes is the overarching question. Using *überhaupt*, a speaker can make overt that she is undertaking a move to a higher QUD, thus making the deviation from the most expected direct answer less burdensome for the hearer.

5 Other roadsigns

In this paper, I explained how language resources such as intonation and discourse particles can act as pragmatic roadsigns that overtly signal specific moves in conversation. As we saw in section 4, this is especially useful when the move is unexpected or marked. I presented the framework of S-trees, and showed that it is both precise and flexible enough to capture the particles *noch* and *überhaupt*, as well as their interaction with presuppositions.

In German, and crosslinguistically, there are many other roadsigns that may provide information about the structure of discourse and the QUDs that are being navigated. For one, the particle *überhaupt* has a number of equivalents and near-equivalents in other languages (Migron, 2005a; Migron, 2005b). This suggests that unifying the various uses of *überhaupt* using QUDs was fruitful, and that other languages have resources to signal a move to a higher QUD. In (Rojas-Esponda, To appear b), I argue that the particle *doch* signals the raising of a previously settled issue. It thus goes against the expectation that questions whose answers are known will not be brought up again (see the maxims of *inquisitive sincerity* or *interactive sincerity* in Groenendijk and Roelofsen (2009) and Coppock and Brochhagen (2013), respectively). The analysis of German *ja* and St'á'imcets *qa7* by Kratzer and Matthewson (Kratzer and Matthewson, 2009) is keyed into a related idea: They analyze ja(p) and qa7(p) as signaling that the question of whether or not p is not currently considered on the table. For Japanese, Davis (2009) convincingly argues that the particle *yo* signals the resolution of the addressee's decision problem (See also McCready (2006)). The overarching idea is that languages have resources, such as intonation and discourse particles, that can help interlocutors coordinate and align their views of the conversation. They might signal a change in the QUD, a move to a higher QUD, whether or not a QUD is considered on or off the table, and whether a QUD has been resolved, among other things. Making precise how this works could give new insights for the view of language as interaction and negotiation (Clark, 1996; Parikh, 2001; Stone and Thomason, 2003; Stone et al., 2007).

Acknowledgements

For helpful discussions, I thank Cleo Condoravdi, Donka Farkas, Dan Lassiter and the Stanford Semantics and Pragmatics group. Special thanks to Chris Potts.

References

Jan Anderssen. 2006. Generalized domain widening *überhaupt*. In Donald Baumer, David Montero, and Michael Scanlon, editors, *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pages 58–66, Somerville, MA. Cascadilla Proceedings Project.

David Beaver and Brady Zack Clark. 2008. *Sense and Sensitivity: How Focus Determines Meaning*. Wiley-Blackwell, Oxford.

Daniel Büring. 1999. Topic. In Peter Bosch and Rob van der Sandt, editors, *Focus — Linguistic, Cognitive, and Computational Perspectives*, pages 142–165. Cambridge University Press, Cambridge.

Daniel Büring. 2003. On D-trees, beans, and B-accents. *Linguistics and Philosophy*, 26(5):511–545.

Robin Clark and Prashant Parikh. 2007. Game theory and discourse anaphora. *Journal of Logic, Language and Information*, 16(3):265–282.

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

Elizabeth Coppock and Thomas Brochhagen. 2013. Raising and resolving issues with scalar modifiers. *Semantics and Pragmatics*, 6(3):1–57, August.

Christopher Davis. 2009. Decisions, dynamics, and the Japanese particle *yo*. *Journal of semantics*, 4(26):329–366.

Regine Eckardt. 2007. Navigating in question answer discourse. *Interfaces and Interface Conditions*, 6:77.

Bart Geurts and Rob van der Sandt. 2004. Interpreting focus. *THEORETICAL LINGUISTICS-BERLIN AND NEW YORK-*, 30:1–44.

Jonathan Ginzburg. 1995a. Resolving questions, part I. *Linguistics and Philosophy*, 18(5):459–527.

Jonathan Ginzburg. 1995b. Resolving questions, part II. *Linguistics and Philosophy*, 18(6):567–609.

Jeroen Groenendijk and Floris Roelofsen. 2009. Inquisitive semantics and pragmatics. In J.M. Larrazabal and L. Zubeldia, editors, *Meaning, Content, and Argument: Proceedings of the ILCLI International Workshop on Semantics, Pragmatics, and Rhetoric*, Amsterdam. ILLC.

Jeroen Groenendijk and Martin Stokhof. 1982. Semantic analysis of wh-complements. *Linguistics and Philosophy*, 5(2):175–233.

Jeroen Groenendijk and Martin Stokhof. 1984. *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Jeroen Groenendijk. 1999. The logic of interrogation: Classical version. In Tanya Matthews and Devon Strolovitch, editors, *SALT IX: Semantics and linguistic theory*, pages 109–126. Cornell University Press, Ithaca.

Ray S. Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.

Ekkehard König. 1983. Polysemie, Polaritätskontexte und überhaupt. In Harald Weydt, editor, *Partikeln und Interaktion*, Reihe germanistische Linguistik 44, pages 160–171. Niemeyer, Tübingen.

Angelika Kratzer and Lisa Matthewson. 2009. Anatomy of two discourse particles. Abstract for SULA 5.

David Lewis. 1988. Relevant implication. *Theoria*, 54(3):161–174.

Sophia Alexandra Malamud. 2006. *Semantics and pragmatics of arbitrariness*. Ph.D. thesis, University of Pennsylvania Dissertation.

Eric McCready. 2006. Japanese *yo*: Its semantics and pragmatics. *Sprache und Datenverarbeitung*, (30):25–34.

Hagit Migron. 2005a. Expressions similar to Hebrew 'bixlal'. *The Linguist List*, 16(268).

Hagit Migron. 2005b. Expressions similar to Hebrew 'bixlal'. *The Linguist List*, 16(999).

Prashant Parikh. 2001. *The Use of Language*. CSLI, Stanford, CA.

Craige Roberts. 1996. Information structure: Towards an integrated formal theory of pragmatics. In Andreas Kathol Jae Hak Yoon, editor, *OSU Working Papers in Linguistics*, pages 91–136. The Ohio State University Department of Linguistics, Columbus, OH.

Craige Roberts. 1998. Focus, the flow of information, and universal grammar. *SYNTAX AND SEMANTICS, VOL 29*, 29:109–160.

Craige Roberts. 2003. Uniqueness in definite noun phrases. *Linguistics and Philosophy*, (26):287–350.

Craige Roberts. 2004. Context in dynamic interpretation. In L.R. Horn and G. Ward, editors, *The Handbook of Pragmatics*, pages 197–220. Blackwell, Oxford.

Tania Rojas-Esponda. To appear a. A discourse model for überhaupt. *Semantics and Pragmatics*.

Tania Rojas-Esponda. To appear b. A QUD account of the German particle doch. *Sinn und Bedeutung 18*.

Mats Rooth. 1985. *Association with Focus*. Ph.D. thesis, UMass Amherst. Revised 1998.

Anders Schoubye. 2009. Descriptions, truth value intuitions, and questions. *Linguistics and Philosophy*, 32(6):583–617.

Matthew Stone and Richmond Thomason. 2003. Coordinating understanding and generation in an abductive approach to interpretation. In *Proceedings of DiaBruck 2003*. Department of Computational Linguistics, Universität des Saarlandes.

Matthew Stone, Richmond Thomason, and David DeVault. 2007. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. To appear in Donna K. Byron; Craige Roberts; and Scott Schwenter, *Presupposition Accommodation*.

Josef Taglicht. 1984. *Message and Emphasis: On Focus and Scope in English*. English language series 15. Longman, London.

Richmond Thomason, Matthew Stone, and David DeVault. 2006. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. *Presupposition Accommodation. Ohio State Pragmatics Initiative*.

Arnim von Stechow. 1981. Topic, focus, and local relevance. In Wolfgang Klein and Willem Levelt, editors, *Crossing the Boundaries in Linguistics*, pages 95–130. Reidel, Dordrecht.

Arjen Zondervan, Luisa Meroni, and Andrea Gualmini. 2008. Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In *Proceedings of SALT*, volume 18, pages 765–777.

Incomplete Answers and the Rise-Fall-Rise Contour

Wagner, Michael, Elise McClay, & Lauren Mak
McGill Linguistics / 1085 Dr. Penfield, Montreal, QC H3A 1A7, Canada
Corresponding Address: `chael@mcgill.ca`

Abstract

The English rise-fall-rise contour has played an important role in the development of theories of intonational meaning, but there have been only few experimental studies testing their predictions. This paper reports on production and perception experiments which investigate the idea that the RFR is an intonational tune that conveys that the current assertion is not a complete answer to the question under discussion.

1 Introduction

Consider the following sentence, uttered with main prominence on *all*. Jackendoff (1972) describes an intonational realization which involves a rising accent on *all*, followed by a fall, and then a final rise at the end of the utterance (in the following indicated by '.../'):

(1) ALL of the men didn't go.../
L H* L- H%

We will refer to this intonation as the rise-fall-rise intonation, or the RFR, following much work in the literature on the topic (Hirschberg and Ward, 1992). It is important to distinguish between the RFR and other rising intonations such as the yes/no question rise, the continuation rise, the incredulity contour, and the contradiction contour (Goodhue et al., 2013). There are two types of analysis of the nature of this intonation. One view takes the RFR contour to be the reflex of a special pitch accent on the word in focus (in this case *all*); the other takes the RFR contour as a sentence-level intonational tune, similar to the declarative contour or the rise typically used in yes/no-questions.[1]

In the following, we first review two past analyses of the RFR (as a pitch accent, and as an intonational tune), and attempts clearly disambiguate them, and then propose a new analysis based on the idea that the RFR signals that an assertion is only a incomplete answer to the question under discussion. A production experiment was conducted and shows that indeed the RFR is preferred when a speaker intends to convey a partial answer, and only rarely used when a complete answer is conveyed. This is to our knowledge the first time that it was shown in a production task that there are contexts in which the RFR is the preferred contour. Two perception experiments try to further elucide the precise pragmatic import of the contour. The final section, §6 discusses the effectiveness of these experiments and their ultimate conclusion.

1.1 RFR as Pitch Accent

The pitch accent analysis presented in Jackendoff (1972) is that both the rise and following fall-rise are due to a special kind of contrastive pitch accent on the word *all*.[2] Jackendoff (1972) calls this accent *background accent* or B-accent, assuming that it is usually placed on discourse-old information. Because it is the last word carrying the accent in the case of (1), the second rise associated with the accent is realized at the sentence end.

An updated analysis of this kind was proposed

---

[1] In this paper we intend the label RFR to be theory neutral, even though the term is from the literature that treats the RFR as a sentence-tune.

[2] Bolinger (1958) also proposes an analysis using accents, positing a B-accent which marks 'connectedness' and 'incompleteness.' However, the proposed accent classification in Bolinger's paper cannot straightforwardly map to Jackendoff's. Jackendoff's B-accent corresponds to a version of Bolinger's 'Accent A' when there is no further accent following, but in the Jackendoff examples in which an A-accent follows the B-accent (not discussed here), Bolinger would categorize the first as an instance of his 'Accent B'. Wagner (2012) argues that Jackendoff's analysis conflates two distinct types of accent: 1) a non-terminal B-accent as a continuation rise which is unrelated to the RFR (similar to Bolinger), and 2) a terminal instance like in (1) of a sentence level RFR-contour (different from both Jackendoff and Bolinger).

in more recent work on contrastive topics (Büring, 1997; Büring, 2003), although under very different assumptions with respect to the semantic meaning and the pragmatic import of the B-accent. Büring's theory is formalized using alternative-semantics (Rooth, 1992), which assumes that each proposition comes with a set of alternative propositions which can play into the overall meaning of the sentence through their interaction with focus and topic operators. Büring's proposal extends this analysis to allow for more complex alternatives, and argues that an utterance can also evoke a set of alternative questions, that is, a set of a alternative sets propositions. The idea for the analysis of contrastive topics is then that an utterance that includes a B-accent triggers the conventional implicature that one of these alternative questions still remains open or 'disputable' after the context has been updated with the contribution of the assertion of the current utterance. We can summarize Büring's insight about the meaning of the RFR as follows:

(2) Disputability Claim
An utterance involving an RFR must leave an answer to a salient alternative question disputable.

1.2 RFR as Intonational Tune

The second type of approach views the RFR as a sentence-level intonational tune, and is more in line with early descriptions of the contour (Pike, 1945; O'Connor and Arnold, 1961). The idea is that, in principle, the RFR can be 'draped' over any utterance independent of the presence or location of a contrast, in contrast to Büring's and Jackendoff's analysis, which both view their B-accent as necessarily evoking contrastive alternatives to the constituent it is placed on. Of course, even in the tune-analysis one might expect that the meaning of RFR could interact with the contribution that a contrastive emphasis makes, if there is one.

Under this view, the RFR-tune has the effect that the last pitch accent of the utterance is realized with a rising accent, which is immediately followed by a fall, and then a final rise is realized at the end of the utterance. In other words, this analysis also makes claims about the presence of a special pitch accent, but views this to be part of the sentence-level contour. This is parallel to other sentence-level tunes, such as the rising intonation observed in questions, which usually is paired with low pitch accents earlier in the utterance, in contrast to the declarative tune which comes with high pitch accents.[3]

Liberman and Sag (1974) argues in favor of this view of RFR as a sentence-tune and against Jackendoff's account in terms of a special contrastive accent.[4] Ward and Hirschberg (1985) come to a similar conclusion, and were the first to make a precise proposal of how to characterize the pragmatic import of RFR. According to their analysis, the RFR conveys speaker uncertainty:

(3) The Uncertainty Claim
The RFR conveys uncertainty with respect to a scale: A speaker conveys uncertainty about whether to evoke the scale, about which scale to choose, or whether the choice of value from the scale is correct.

This analysis is similar to Büring's in that if there remains uncertainty, this plausibly means that there is still an unresolved and hence disputable issue that remains open. It differs from Büring's analysis in that there could be uncertainty with respect to the present assertion that carries the RFR, by virtue of the last clause in (3). Büring's analysis does not predict any uncertainty about the present assertion.

It is important to point out, however, that this difference in the pragmatic/semantic analysis between the two views is only loosely linked to the difference in whether or not RFR is treated as a pitch accent or a tune. In other words, one could imagine a Büringian tune-analysis and a Ward & Hirschbergian pitch accent analysis. Constant (2012) and Wagner (2012), for example, posit a tune-analyses couched in alternative semantics, and proposes a meaning for the tune that is very similar to Büring's disputability implicature. For the remainder of this paper we focus on the *semantic/pragmatic differences* between accounts and leave a discussion of the tune-vs.-accent issue for another occasion.

---

[3] Pierrehumbert and Hirschberg (1990), however, argue that pitch accent type and boundary tone (delcarative vs. question rise) are freely combinable—we have some reservations about this claim but will not elaborate on this here.

[4] Arguably Liberman and Sag (1974) conflate the RFR with a different contour, the Contradiction Contour (not discussed here). A running theme in the literature on tunes is that different authors assume different taxonomies of intonational tunes, which makes a direct comparison between proposals difficult.

1.3 Disambiguation?

An important difference between the analysis in Ward and Hirschberg (1985) on the one hand and the analyses of Büring, Constant and Jackendoff on the other is that in the latter, the use of the special intonation is predicted to have the effect that the universal quantifier takes scope below negation. The sentence in (1) should then be paraphrasable as follows:

(4) Not ALL of the men went.../

Büring's analysis offers an elegant account of this (purported) disambiguation. Suppose that (1) didn't involve inverse scope, and instead could be paraphrased as follows:

(5) NONE of the men went.../

Under this reading, the set of alternative questions would look as follows in this analysis:

(6) a. Did all of them go?
b. Did some of them go?

Clearly, the assertive content of (5) already provides an answer to all of these questions, and hence the disputability implicature should be infelicitous. Under the inverse scope reading, the alternatives are the following:

(7) a. Didn't all of them go?
b. Didn't some of them go?

Here, the assertion of (4) leaves open the possibility that none of them went, and only resolves the first question; hence this reading is compatible with the RFR.

Constant (2012) incorporates Büring's insight into a tune-analysis of the RFR. Under this analysis, the RFR obligatorily associates with focus and requires that all propositional alternatives remain unresolved. Asserting (5), with focus on *all*, would then be incompatible with the implicature of the RFR since it resolves all propositional alternatives:

(8) a. None of them went.
b. Some of them went.

Both Büring (1997) and Constant (2012) predict that the RFR should be infelicitous on utterances which resolve all alternatives, even if the technical explanations slightly differ. In other words, an utterance like (5) in which the assertion entails the falseness of all alternatives should be infelicitous with the RFR contour, and this, under these analyses, is the source of disambiguation in (1).

The disambiguation claim has not, however, gone unchallenged. For example, Ward and Hirschberg (1985) argue that context can disambiguate the sentence one way or another, orthogonal to which intonation is used. For more recent evidence bearing on this question see Syrett et al. (2013). While this paper will not directly test Jackendoff's original disambiguation claim, we will see evidence bearing on the predictions of Büring's and Constant's view that alternative-excluding contexts should be incompatible with the RFR.

2 RFR and Incomplete Answers

While we conducted this study to establish some of the basic data points relevant for these prior studies and test some of their diverging predictions, the design of the experiment was motivated further by an analysis that unifies insights from prior analyses but is slightly different in the precise content it attributes to RFR.

Each prior analysis of the contour has been based on one particular use of the contour, which was then argued to generalize to other uses, and ours is no exception to this pattern. The use of RFR we take as a starting point is this:

(9) **Q:** Who solved the problem?
a. **A:** JOHN did.
b. **A:** JOHN did.../

When A wants to provide a partial answer to the question Q raised, then she might use the RFR to signal this. The analysis of the RFR we propose is the following:

(10) **RFR** (p): The speaker asserts p but considers it to be only an incomplete answer to the question under discussion.

An answer with a declarative fall as in (9a) comes with the implicature that no one else but John solved the problem. This is the reading of the sentence that ensues when we treat the answer to be an exhaustive question under the discussion. The use of the RFR in (9b) preempts this implicature.

The analysis is essentially that embodied in Büring's disputability claim, which also requires a question to remain unresolved, but differs in one aspect: The remaining disputable question is always the one the speaker takes to be the question under discussion (QUD). We already know from Büring (2003) that what a speaker takes to be the current question under discussion does not always have to be the question immediately asked before, so some apparent counterexamples are compatible with this view, if we allow the speaker to perform discourse moves that don't just obey the QUD made salient by prior discourse. For example, a speaker might use the RFR to signal that she considers another question related to the present one by forming part of a super-question (in our analysis the unresolved QUD) to be salient:

(11) In a context where it matters how well several people did on a specific problem...
Q: Did John solve the problem?
A: JOHN did.../ (But SALLY didn't!)

The immediate question under discussion asks about one particular individual, but the intonation in the answer reflects the fact that speaker A considers a broader question, namely *Who solved the problem*, or maybe *Did John and Sally solve the problem?*

Our analysis of RFR as a contour signalling a partial contour simplifies Büring's in that it does not require the complex topic-semantic value for utterances that Büring's analysis relies on, and uses a 'flatter' representation that only involves the meaning of utterances and alternatives to that utterance. An argument against topic-semantic values and in favor of 'flatter' meanings is given Wagner (2012). This fits with the observation in Wolter (2003) and Constant (2012) that partial answer are one typical environment for the RFR.

Our analysis differs from the analysis in (Ward and Hirschberg, 1985) in that it does not attribute uncertainty to the meaning of the contour itself. Rather the assumption is that that uncertainty inferences are a conversational implicature that result from choosing the RFR over a declarative contour. Not providing a complete answer is felicitous if the speaker does not know the complete answer, but would often be uncooperative otherwise. There may be uses of partial answers though that are not uncooperative even when the speaker does know the complete answer, and we would predict in those cases no inference about uncertainty to ensue. One such use of the RFR is arguably when it is used to insinuate something and there is either a taboo or some other reason why the speaker does not want to state something explicitly. We will return to this point below.

This paper reports on a series of experiments in which we first test whether speakers indeed use the RFR to convey an incomplete answer. First, a production experiment tests whether speakers actually use the RFR when providing incomplete answers, and whether they avoid it when they provide a complete answer. Second, two perception experiments are used to test the claims about the meaning of the contour further.

3 Experiment 1: Production

What are the odds that a speaker uses the RFR in a situation where, based on a particular analysis of what the RFR contributes pragmatically, the RFR would seem like a good option? To our knowledge, no previous account of the RFR has tried to establish this empirically, maybe because of a sense that the intonation is elusive and even rare when the conditions of its use are optimally met.

What are the odds that the RFR is not used in a context that a particular account predicts to be incompatible with its use? Even for this question there has been very little experimentation, since most of the semantic work is based on impressionistic intuitions alone.

A production experiment was conducted to test whether the RFR is used in partial answers and not in complete answers. A second goal of this study was to collect a mini-corpus of utterances in which the RFR and other tunes were used by speakers without prior training on intonational tunes or priming that certain tunes were even a possible choice. These utterances will then be used in perception studies to elucidate what the contours were taken to mean when heard in or out of context.

3.1 Methodology

Participants were asked to respond to a prerecorded question played to them over headphones. They were unaware that the experiment was about sentence intonation. Prior to recording the dialogue, participants read the script of the dialogue, which included a set of 'stage directions' with respect what their intentions were supposed to be with their assertion. The crucial manipulation was that they were effectively told whether their answer is intended as a complete answer or not. An example dialogue with a complete answer, predicted to disfavor use of the RFR:

(12) **Q:** Is Bill coming to the party?
[You wish to convey that you know Bill will be coming to the party.]
A: Bill is coming.

The answer completely resolves the QUD. An partial-answer tune would only be motivated here if A actually considers a broader question, maybe *Who is coming to the party* or *Are Bill and some others coming to the party?*. But expanding the discourse to a broader question is not motivated by the context, hence we expect a lower rating for appropriateness.

An example with a context that is predicted to favor the use of the RFR:

(13) **Q:** Is either Bill or Susan coming to the party?
[You know for sure Bill is coming, but you wish to convey that you are not sure whether Susan is coming as well.]
Is Bill coming to the party?
A: Bill is coming.

Here, the question leaves one part of the QUD unresolved, and this is also made explicit in our stage directions. Participants were asked to read the question, context, and answer silently until comfortable with the material, and then read out only the reply as if in a normal conversation once prompted with the pre-recorded question. Twenty native speakers of North American English were tested. The experiment had a latin square design, such that each participant took part in 8 dialogues, 4 from each condition, in pseudo-random order such that repetitions of condition and item were minimized; this yielded a total of 160 utterances. The experiments were run using a set of Matlab scripts. The data was then acoustically analyzed (reported briefly here), as well as annotated by a trained Research Assistant (RA) for which contour was used:

(14)
a. RFR
b. Question Rise
c. Falling Contour
d. Unclear/Other

3.2 Results

The annotation summarized in Figure 1 shows that the contextual manipulation was successful in creating both situations in which participants were likely to use the RFR contour, and ones in which they were unlikely to use it (the data reported on here only includes utterances from the first three categories in (14), which comprised more than 90% of all trials). In the dialogues that involved answers that were necessarily complete answers to the QUD, a falling declarative contour was used more than 83% of the time and an RFR contour less than 12% of the time. In the dialogues that involved answers that were compatible with being partial answers, and in which stage directions had made it clear that only a partial answer was intended, the RFR contour was used more than 65% of the time, and the declarative contour only 24% of the time.

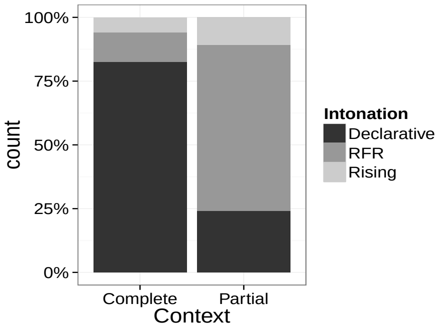


Figure 1: Experiment 1. Percent of Declarative, RFR, and Rising Intonation in Annotation

We tested that the difference was indeed statistically significant by looking at the subset of data that only involved Declarative or RFR contours, and fitting a mixed effects logistic regression model with RFR (presence or absence) as the dependent variable, Context as a fixed factor, and mixed effects for Item and Participant that included a random slope for Context. The contribution of Context was highly siginificant at $p < 0.001$.

As way to check that there were indeed systematic acoustic differences between the different contours, we report here about a single measure, the maximum pitch in the final quadrant of the final word of the utterance. We fit a linear mixed model with this dependent measure and Intonation (levels: Declarative, RFR, Rising) as fixed factor and random effects for participant and item, including random slopes for Intonation. We found highly significant differences between Declarative vs. RFR (t>2.8), such that there was a higher final pitch in the cases where the annotator labeled a contour as RFR. We will not explore the acoustics in more detail in this paper.

3.3 Discussion

Experiment 1 is, to our knowledge, the first production experiment in which it was shown that in certain contexts and while trying to convey a certain meaning, speakers are likely to use the RFR contour. In fact, in these contexts the RFR contour is more likely than any other contour, and it is rare in other contexts.

Why was the RFR ever used in complete-answer contexts? One possibility is that our annotations conflated different prosodies (e.g., the incredulity or contradiction contours also involve a fall rise). Another possibility is that our manipulation wasn't successful in requiring a complete answer all the time. A disproportionate fraction of the RFR used in the complete contexts were due to two particular items, making this option seem likely. Finally, over the course of the experiment, participants might just have paid less attention and repeated the contour they used on the last trial. In order to establish whether the contour has indeed the pragmatic import we assume, and also to get at the diverging predictions of different theories, we ran two perception experiments.

4 Experiment 2: Appropriateness Rating

Is the RFR contour really dispreferred in complete-answer contexts? Our assumption about what the RFR conveys suggests that it should be. In order to answer this question, we ran a perception study in which listeners had to rate how natural a response sounds given the dialogue context. We used utterances from six speakers from experiment 1–these utterances were sampled from 4 of the original 8 items, and played either in the original context or in opposite one. The items we sampled the productions from were those in which the context itself, even without the stage directions, makes a partial answer interpretation of the response unlikely.

4.1 Methodology

Participants were presented with the audio recordings as described, either matching utterances from Experiment 1 to their appropriate context, or playing audio that mismatched the condition of the context and response. Participants were then asked to rate the response on a scale from 1-7, where *higher numbers* indicated the utterance was *more appropriate* as a response to the given question. Here is an example with predictions:

(15) Complete-Answer Context:
    a. Q: Is Bill coming to the party?
       A: Bill is coming. (Declarative)
    b. Q: Is Bill coming to the party?
       ? A: Bill is coming.../ (RFR)

(16) Partial-Answer Context:
    a. Q: Is either Bill or Susan coming to the party?
       A: Bill is coming. (Declarative)
    b. Q: Is either Bill or Susan coming to the party?
       A: Bill is coming.../ (RFR)

It is of course also possible to use the RFR in a complete-answer context. This could mean that speaker A wants to convey that she is considering a broader question than the one asked, or that she wants to move the discourse to such a broader QUD. But, since this is not motivated by the context, we expect the rating for an RFR contour answer to be lower in response to a complete-answer context.

4.2 Results

Figure 2 summarizes the results. The top panel shows the appropriateness rating by context, depending on which original context the utterance was recorded in.

As predicted, the only case in which the appropriateness seems lower is the one where an utterance was recorded in a partial context (those in which the RFR was most commonly used) and was then played in the complete context (the context in which the RFR is not expected to occur). If an utterance was recorded in the *complete* context and is then played back in the *partial* context, there is a much smaller difference, or none at all.

We analyzed the data using a mixed model regression with the original context and the new context, and their interaction as random effects, and random effects for participant and item, that included slopes for the interaction. The main effects of original and new context were not significant,

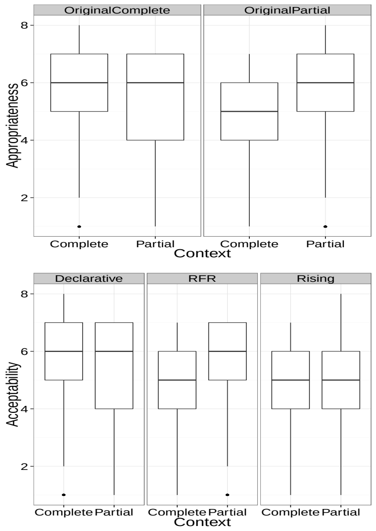


Figure 2: Experiment 2: Acceptability by original context (top) and intonation (bottom). Question: 'How appropriate do you find the answer is pronounced given the question? (From 1=completely inappropriate to 7=completely appropriate)'

although the effect of original context approaches signficance ($t > 1.0$). The interaction between original and new context was highly significant ($t > 3.1$), as predicted.

These results show that the answer's original context mattered in determining which context listeners deemed the utterance more appropriate. But they don't show yet *why* they were more or less appropriate. Our hypothesis was that the observed differences were due to the RFR contour. The bottom panel in Figure 2 shows the results by intonation. We see, as expected, that the RFR is less suitable for the complete context, but the declarative contour is suitable for both contexts, while rising (question) intonation is rated as less suitable overall in both types of context.

Again, we fitted a mixed model, this time with intonation and context, and their interaction as fixed effects, and random effects with slopes for participant and items. There was a main effect of context such that utterances played in Complete contexts were deemed more appropriate ($t > 2.1$), and crucially there was a highly significant interaction between context and intonation: The difference between Declarative and RFR was different in Complete vs. Partial contexts ($t > 2.8$).

4.3 Discussion

The results show that using the RFR contour is more compatible with contexts in which the current assertion can be taken to be an incomplete answer. In contexts in which the assertion seems to be a complete answer to the question under discussion, the RFR contour does not appear to be completely infelicitous, in contrast to the claim in Constant (2012) that the RFR is incompatible with uses in utterances for which all alternatives are resolved (i.e., in complete answers). Since Constant (2012) assumes obligatory association with focus, alternatives that are not structurally related (such as broader question that are not part of the formal set of alternatives) are not available in the interpretation of RFR.

We do not assume association with focus with the RFR to be obligatory or even necessary. Therefore is expected that a speaker might use the RFR to convey that she considers a super-question other than the question in the immediate context to be the QUD, in which case the provided response is indeed a partial answer. For example, in the following dialogue, A answers the immediate question under discussion, but the RFR indicates that A assumes the relevance of a larger question. The RFR then conveys that A is not in a position to answer that, as indicated by the continuation after the first sentence:

(17) Q: Is Bill coming to the party?
A: Bill is coming.../ But I'm not sure whether anyone else is...

In order to get more specific information about the meaning of the contour, we ran a second experiment, in which we asked listeners more directly about what they think an utterance communicates.

5 Experiment 3: Guessing Intentions

Suppose we play utterances out of context: What do listeners infer about the intended meaning? We ran an experiment in which we directly asked participants what they think a speaker wanted to convey, addressing three separate qualities. If the RFR really conveys uncertainty, what does it convey uncertainty about? Is it 1) about the confidence that the proposition that is asserted itself is true, or 2) about the fit of the assertion into the current discourse? We also examined a third option, 3) that the RFR is being used to insinuate something above and beyond the asserted content (regardless of certainty).

5.1 Methodology

We used the same materials as in Exp. 2. This time, listeners only heard the “answer” recordings of the question-answer pairs, thus removing the influence of the context on the interpretation of the contour. Rather than simply ranking how appropriate the sentence was out-of-context, this time participants rated the utterances on three separate Likert-scales that directly addressed what the listeners thought were the intentions of the speaker.

5.2 Results

Figure 3 summarizes how confident listeners thought speakers were in the asserted content. There was a non-significant trend that utterances originally produced in *Partial* contexts were rated as less confident about the asserted content (Mixed model analysis, $t = 1.81$). When looking at Intonation, however, the Declarative contour differed significantly from the RFR contour ($t = 2.59$). The rising contour did not differ from the RFR in this respect. Only the approach in Ward and Hirschberg (1985) expects the fact that the RFR contour was taken by the listeners to convey that speakers were less confident about the asserted content–it comes as a surprise for the approaches in (Jackendoff, 1972; Büring, 1997; Constant, 2012; Wagner, 2012).

Figure 4 shows how confident listeners thought speakers were that their utterance fit the discourse context. Uncertainty about whether an assertion is relevant to the context is another dimension about uncertainty which was attributed to the RFR by previous approaches, particularly the one in Ward and Hirschberg (1985). There was a non-significant trend that utterances originally produced in *Partial* contexts were rated as conveying speaker uncertainty that the assertion is context-appropriate (Mixed model analysis, $t = 1.92$).

Again, when looking at Intonation, the Declarative contour differed significantly from the RFR contour ($t = 2.8$), such that listeners thought speakers were less confident about fit, which the Rising intonation did not ($t = 0.89$).

Figure 5 shows how likely listeners thought it was that speakers were insinuating something with their answer above and beyond the literal meaning they were conveying. Utterances originally produced in *Partial* contexts were rated as much more likely to insinuate something. The RFR contours differed from both Declarative ($t = -2.4$) and Rising contour ($t == 2.6$) in this regard. This is what is predicted under the present analysis in the absence of a context question, since the RFR indicates that some issue still remains open.

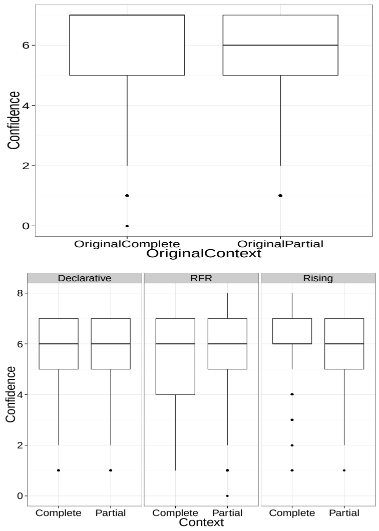


Figure 3: Experiment 3, Question 1: Confidence in Asserted Content. Question: ‘How confident do you think the speaker is about the literal statement she/he is making? (Between 1=not confident at all, and 7=completely confident)’

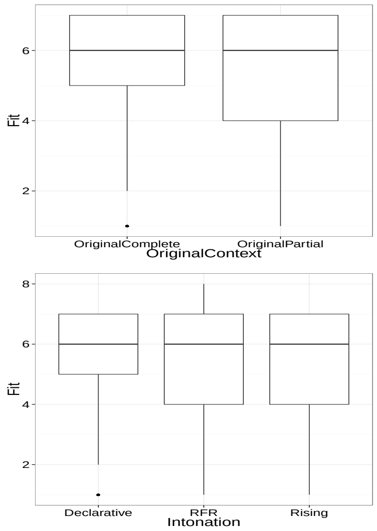


Figure 4: Experiment 3, Question 2: Confidence of Fit into Context. Question: ‘How confident do you think is the speaker about whether or not the answer is relevant for the present discussion? (Between 1=not confident at all and 7=completely confident)’

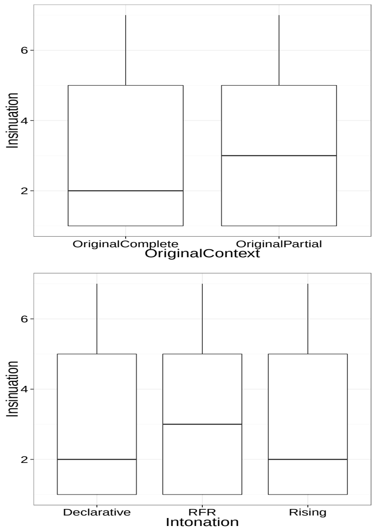


Figure 5: Experiment 3, Question 3: Likelihood of Insinuation. Question: 'How likely do think is it that the speaker tries to insinuate something apart from what she/he is actually saying? (Between 1=not likely at all and 7=very likely)'

5.3 Discussion

The results are compatible with the present analysis, and raise some questions for the alternative accounts. The least accounted-for effect is that we did not expect the RFR to convey uncertainty about the present assertion, and neither do most analyses of the RFR contour. One striking result is that the RFR is overwhelmingly taken by listeners to indicate that the utterance is meant to insinuate something above and beyond what is literally asserted by the speaker. That the RFR can be used to insinuate non-asserted content is compatible with all analyses considered here.

6 Conclusion

We proposed that using the RFR contour allows speakers to encode that they consider their assertion to be an *incomplete answer* to the question under discussion. We showed that based on this hypothesis, we can create contexts in which speakers are highly likely or unlikely to use the RFR contour, namely by asking them to convey partial or complete answers respectively. This is, as far as we know, the first time that it was shown that the RFR is the preferred (i.e., most frequent) intonational contour used in some contexts. Showing that a contour is actually used in a production task in which no reference is made to the participants that the experiment is about intonation is important since it provides real evidence that the contour is systematically used at least under certain circumstances. Perception experiments that try to get at this question have the problem that by providing examples of a contour they invite the listeners to consider they make the existence of the contour salient, and this might bias results when listeners compare their felicity to other contours.

Two perception experiments complemented the production evidence and provided further insights into the pragmatic import of the contour. The RFR was rated less felicitous in a context which favors a complete answer, but it is far from being infelicitous. In our analysis, in these cases a broader question has to be accommodated to make sense of the fact that the contour signals that the speaker considers the response to be a partial answer. Infelicity would be expected based on the claim in Constant (2012) that the RFR obligatorily associates with focus, and can only express uncertainty about alternatives structurally related to the asserted one. This precludes the possibility that a broader issue outside of the alternative set evoked by the assertion can be raised by the speaker using the RFR.

The contour is compatible with cases with in which it is not obvious that a scale really plays a role, which raises some questions with respect to the proposal in Ward and Hirschberg (1985). On the other hand, the contour also seems compatible with conveying uncertainty about the present assertion—this is compatible with Ward and Hirschberg (1985)'s account, but unexpected the other approaches discussed in this paper, including our own proposal. More experimental work is necessary to better understand this contour and the conditions on its use.

A very clear result of our perceptual studies is that out of context, the RFR is taken by listeners to indicate that the utterance is meant to insinuate something above and beyond what is literally asserted by the speaker. Our analysis makes the following prediction about the sense of insinuation that the RFR conveys: When used in a context where it is obvious why the provided answer might be incomplete, the only insinuation conveyed should be that the speaker is not ready to provide a complete answer; however, when the asserted content appears to answer the question under discussion completely, then indicating that the answer is incomplete should signal that *the* *speaker in fact considers something else to be the real issue*, something that is still open and unresolved.[5]

A direct test of this prediction would be to ask the question about insinuation from Experiment 3 in a perception experiment that plays both context and answers to listeners, a prediction that we haven't tested yet. The sense of insinuation should be greater in complete contexts. If true, it would mean that the lower appropriateness rating of the RFR in complete contexts in Experiment 2 may not result from an inherent incompatibility between the RFR and such contexts, but rather from listeners not being able to guess what was being insinuated, and hence judging the utterances as less appropriate.

Acknowledgments

This paper started out as class project conducted by the second author in a class on laboratory linguistics taught by the first author in 2010. Thanks to everyone in prosody.lab for their help in running these studies, in particular to Erin Olson and Aron Hirsch. Thanks also to Noah Constant and Dan Goodhue for comments and discussions about the RFR. This research was funded by grants from the FQRSC and from SSHRC, as well as from the Canada Research Chair Program.

References

D. Bolinger. 1958. Stress and information. *American Speech*, 33:5–20.

Daniel Büring. 1997. *The Meaning of Topic and Focus: The 59th Street Bridge Accent*. Routledge Studies in German Linguistics. Routledge, London.

Daniel Büring. 2003. On D-trees, beans, and B-accents. *Linguistics and Philosophy*, 26(5):511–545.

Noah Constant. 2012. English rise-fall-rise: A study in the semantics and pragmatics of intonation. *Linguistics and Philosophy*, 35(5):407–442.

Dan Goodhue, James Pickett, and Michael Wagner. 2013. English reverse prosody in yes-no responses. In *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*.

Julia Hirschberg and Gregory L. Ward. 1992. The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, 20:241–251.

Ray S. Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, Ma.

Mark Liberman and Ivan Sag. 1974. Prosodic form and discourse function. In *Proceedings of CLS*, volume 10, pages 416–427.

J.D. O'Connor and G.F. Arnold. 1961. *Intonation of Colloquial English*. Longman, London.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan, and Marth E. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT Press, Cambridge, Ma.

K.L. Pike. 1945. *The intonation of American English*. University of Michigan Press, Ann Arbor, MI.

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

Kristen Syrett, Georgia Simon, and Kirsten Nisulab. 2013. Prosodic disambiguation of scopally ambiguous sentences. In *Proceedings of NELS*.

Michael Wagner. 2012. Contrastive topics decomposed. *Semantics & Pragmatics*, 5(8):1–54.

Gregory Ward and Julia Hirschberg. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61(3):747–776.

Lynsay Wolter. 2003. Fall-rise, topic and speaker non-commitment. In *Proceedings of WECOL 2002. University of British Columbia*.

---

[5] Another possible use of the RFR is in answers that insinuate an answer to the question under discussion, but their literal content appears to not address it. Here's an example from Wagner (2012):

(i.) A: Did Mary rob the candy store?
B: She likes chocolate.../

The speaker can use this response to convey that she thinks Mary did rob the candy store or at least thinks it's plausible. It is not clear that our proposal is sufficient to explain such uses of the RFR. In some sense, even if the speaker wants to insinuate that she thinks that Mary did rob it (and hence provide a complete answer), the answer in itself does not provide a complete answer. Our proposal needs to be made more explicit by defining what counts as an incomplete answer. Wagner (2012) argues that the RFR conveys that a salient alternative to the asserted proposition is possibly true. This account seems better suited for this type of example—we defer further discussion to a future occasion.

‘Attention, I’m violating a maxim!’
A unifying account of the final rise

Matthijs Westera
Institute for Logic, Language and Computation
University of Amsterdam
m.westera@uva.nl

Abstract

Declarative sentences that end with a rising pitch in English (among other languages) have many uses. I single out several prominent uses that the literature so far has treated mostly independently. I present a compositional, unifying analysis, where the final rising pitch marks the violation of a conversational maxim, and its steepness indicates the speaker’s emotional activation. Existing theories are reproduced from these basic assumptions. I believe it contributes to a solid theoretical foundation for future work on the semantics and pragmatics of intonation.

1 Introduction

Declarative sentences in English (among various languages) can end with a rising pitch (as defined, very liberally, in section 2.1). This final rise has at least three prominent uses, which I will conveniently name by the Kantian categories (notably used, of course, by Grice (1975)):

1. **Quality reading:** that the speaker is uncertain whether what she is asserting is true;
2. **Quantity reading:** that she is about to say more, or at least knows more, on the present topic, than what she is asserting; and
3. **Relation reading:** that she is uncertain about (how her response relates to) some alternative answer to the question.

These readings are illustrated in the following examples, where ↗ marks the relevant final rises.

(1) **Quality reading:**
  a. A: John has to pick up his sister.
     B: John has a sister↗.
     (Trinh & Crnič, 2011)
  b. A: Guess which colours John likes!
     B: He likes blue↗.

(2) **Quantity reading:** (or ‘list intonation’)
  a. A: Who was at the party?
     B: Mary↗, Bob↗, and Sue.
  b. A: What did you do today?
     B: I sat in on a history class↗.
     I learned about housing prices↗.
     And I watched a cool documentary.
     (Tyler, 2012)

(3) **Relation reading:**
  a. A: Was John at the party?
     B: (Well,) it was raining↗.
  b. A: Of John, Mary and Bob, who came to the party?
     B: (Well,) John was there↗.

This is a remarkable combination of readings. For instance, the Quality and Relation readings suggest that the final rise conveys speaker uncertainty (as has been proposed in the literature, e.g., Gunlogson, 2003; Truckenbrodt, 2006; Ward & Hirschberg, 1992; Constant, 2012), but this is at odds with the Quantity reading. And while the Quantity and Relation readings pertain to what has not (yet) been asserted, the Quality reading pertains to the asserted proposition itself. In addition to this semantic variation, there exist intonational differences between the readings, in particular in the steepness of the rise (as discussed in section 2.1). For these reasons, one might think that to try and give a unified analysis of the three readings would be a misguided and hopeless attempt.

Nevertheless, I will show that a unified analysis is possible. I present a compositional analysis, where the final rise (whether high or low) marks **the violation of a conversational maxim** (hence the Kantian/Gricean labels), and its steepness indicates **the speaker’s emotional activation**. The main burden of my account is carried by a precise formulation of the maxims, which I adopt from the literature. In section 2 I present the main ingredients of my approach. In section 3 I show how it predicts the three readings (and one more) and compare it to existing accounts proposed for each reading in isolation. Section 4 discusses the predictive power of the theory. In section 5 I conclude, and identify directions for future research.

2 Ingredients

2.1 What is a final rise

In the literature, what I call the Quality reading has been assigned primarily to a *high* (or steep) final rise (Gunlogson, 2003), while the Quantity and Relation readings have been assigned to a *low* final rise, with the Relation reading being associated in particular with the entire *rise-fall-rise* contour (Constant, 2012). To give a unified account, we therefore need a very liberal definition of 'final rise', as well as an explanation of the phonological differences between the readings.

I consider as a 'final rise' any contour whose tail (the part after the nuclear stress) is non-falling throughout its end. This is a more liberal notion of 'final rise' than that employed by Gunlogson (2003), who follows Gussenhoven (1983) in requiring that the final pitch is higher than the nuclear accent (a requirement we drop, crucially, because we claim that the final pitch has an indepedent semantic contribution). Gunlogson also excludes contours with bitonal (rising) accents, such as rise-fall-rise. However, I believe that bitonal accents have an independent semantic contribution, one that is orthogonal to our discussion. Although the literature associates the rising accent with the Relation reading, it seems to me that a rising accent is neither necessary nor sufficient for it. That is, (2a,b) can be read with a simplex accent, and, conversely, (1a,b) can be read with a rising accent, i.e., with a rise-fall-rise contour (perhaps conveying extra surprise). Indeed, Ward and Hirschberg (1992) show for the rise-fall-rise contour that a lower rise triggers a Relation reading (their '(scalar) uncertainty'), while a higher rise triggers a Quality reading (their 'incredulity').

Gunlogson's (2003) notion of 'final rise' is already quite liberal, and ignores a lot of variation.[1] The null-hypothesis, I think, is that all variation is due to the stacking of several intonational components, each with its own, independent semantic contribution. Hence, distilling two components to study independently - the final rise and its steepness - is methodologically sound.[2]

2.2 Semantics and pragmatics

Following, e.g., Gussenhoven (1983), I treat the final contour as an independent meaning-carrying component. For concreteness, I assume that its semantic contribution is *non-at issue content* (following, e.g., Ward and Hirschberg (1985); Constant (2012) for rise-fall-rise). I assume that the final rise semantically takes an *expression* as its argument.[3] I assume that, on declarative sentences:

- the final rise conveys that uttering the expression in the present context would violate a conversational maxim;[4] and
- the relative height of the final pitch indicates the speaker's emotional activation.

The first assumption is perhaps novel in its generality (as pertaining to *any* maxim), but certainly not in spirit. For instance, Ward and Hirschberg (1985) already write that 'intuitively, [rise-fall-rise] seems to indicate that a speaker is uncertain about whether his utterance is relevant to the discourse'. The second assumption also appears to go far back, but we base it in particular on Banziger and Scherer (2005), who found specifically that the steepness of a final rise (as well as a final fall) correlates with higher emotional activation.

What the violation of a maxim amounts to depends, of course, on which maxims there are, and what they require. The following set of maxims is generally accepted as the minimal backbone, where the *QUD* is taken to be an explicit or implicit *question under discussion*:

- **Quality**: Only say what you think is true. (Grice, 1975)
- **Quantity**: Give the most informative answer to the QUD that you think is true.

---

[1] Gunlogson (2003) defends this on the grounds that not all phonological distinctions need to be semantically relevant. I disagree, though perhaps only on her use of the word 'semantic'. I think assigning a semantic distinction to every phonological one would simply require a much richer semantics.

[2] Probably more subtle intonational features may disambiguate among the various readings - or non-intonational features, for that matter, such as shrugging one's shoulders (to exclude the Quantity reading) or counting on one's fingers (to trigger it). Discourse particles or hesitation markers (like 'well' in (3)) may provide additional cues.

[3] It takes an expression as its argument, rather than its meaning, because the semantic contribution of the final rise, as argued below, may also pertain to *how* something is said.

[4] The careful formulation 'uttering the expression' is necessary because the maxims pertain not to expressions, but to utterances. Alternatively, one would have to treat the final rise as a *speech act modifier*, a possibility that is left unexplored in the present paper.

(Groenendijk & Stokhof, 1984; Schulz & Van Rooij, 2006)

- **Relation**: Let your utterance, relative to your information state, entail the QUD.[5]
  (cf. Groenendijk & Stokhof, 1984; Roberts, 1996; Westera, 2013)
- **Manner**: Only utter what you think is clear, concise, etcetera. (Grice, 1975)

For the present purposes only the Maxim of Relation will require some further formalisation. This formalisation is postponed to section 3.3.

Although the final rise conveys that uttering the expression would violate a maxim, examples (1) to (3) above do not seem to involve any true non-cooperativity. This is because a speaker may have a good *reason* for violating a maxim, namely that not doing so would have violated another maxim (for instance, left implicit here, that one should at least *try* to make a useful contribution, even if one is uncertain). That is, the kinds of violations that occur when a cooperative speaker uses the final rise are of the Gricean (1975) 'group B'-type, involving a clash between two maxims. Presumably, only those maxim violations have to be marked by a final rise that might otherwise mislead the hearer (cf. Grice's 'silently violating a maxim').

Since Grice (1975) it has been assumed that a violation of the Maxim of Quality is more dramatic (more non-cooperative) than a violation of the Maxim of Quantity or the Maxim of Relation. I assume that the speaker's emotional activation, in the presence of such violations, reflects this. Therefore, if a final rise marks the violation of a maxim, then typically a *high* rise will mark a violation of the Maxim of Quality, while a *low* rise indicates a less dramatic violation, i.e., the Maxim of Quantity or the Maxim of Relation. Note that this predicted correlation is only *typical*, because (i) the relative importance of the maxims may vary across contexts, and (ii) contextual sources of emotional activation can interfere. For instance, when B in (1b) is completely uninterested in the truth of her guess, the rise for a Quality reading is predicted to be less steep than usual; and when a speaker is very excited about the party guests on the list she is reading in (2a), the rises in her list intonation may be much higher than usual.

3 Deriving the readings

3.1 The Quality reading

The Maxim of Quality requires that the speaker thinks that what she says is true. Therefore, if a final rise conveys a violation of the Maxim of Quality, it conveys that the speaker *lacks the belief* that what she says is true. I assume that this lack of belief lies at the core of the Quality reading, illustrated by the examples in (1). Because a violation of Quality is quite dramatic, it is predicted that this reading typically occurs with a high final pitch, as seen in the literature (cf. section 2.1).

On top of this, additional pragmatic reasoning may shape what exactly the Quality reading amounts to. For instance, as mentioned, Ward and Hirschberg (1992) discern an *incredulity* reading for cases like (1a). This can be analysed as an implicature: if B conveys (by means of the final rise) that she lacks the belief that John has a sister, even though A just said so, that might plausibly be because B finds it hard to believe.

An implicature of the Quality reading that has received the most attention in the literature, is the *contextual bias* in favour of the proposition expressed (e.g. Gunlogson, 2003; Truckenbrodt, 2006; Trinh & Crnič, 2011). The following example illustrates this (one of many by Gunlogson):

(4) <u>Windowless room</u>
  a. Is it raining? (OK without evidence)
  b. # Its raining↗. (OK only if the addressee just entered with an umbrella)

Space does not permit a discussion of all approaches to capture this bias. I will discuss only Truckenbrodt's (2006), which is closest to mine.

Truckenbrodt's (2006) account

The main ingredient of Truckenbrodt's account of the final rise is that it indicates the speaker's *lack of belief* in the proposition expressed. In addition, he assumes that in uttering a declarative (whether rising or falling), a speaker conveys her intention to make the expressed proposition common ground. Hence, with the latter assumption, which I am happy to make, my account of the Quality reading amounts exactly to Truckenbrodt's.

With this, Truckenbrodt explains the bias as follows: a speaker who, by uttering a declarative φ, expresses her desire to make φ common ground, implies that she considers it possible that it will be common ground. If, at the same time, with the

---

[5] Groenendijk and Stokhof (1984) require that the utterance entails (an answer to) the question relative to the *hearer's* information state, and Roberts (1996) relative to the *common ground*. Westera (2013) argues that these requirements are too strict.

final rise, she conveys that her information state does not support φ, this implies that she considers it possible that the addressee's information state *will* support it (for otherwise, it would not be possible for φ to become common ground). This explains why, in the absense of evidence that the addressee might know φ, as in (4), a declarative with a final rise is strange (at least for obtaining the Quality reading). Note that in examples (1a) and (1b), the context makes it clear that A should know something about what B says.

Truckenbrodt's account is the most minimalist among existing approaches to the Quality reading of the final rise (for instance, Gunlogson's (2003) account is formulated in terms of *discourse commitments*, and recently Trinh and Crnič (2011) propose that rising declaratives are *second-person speech acts*, a concept that I have some difficulty grasping). To my awareness, Truckenbrodt's account is also empirically adequate.[6] I think that the fact that the core of Truckenbrodt's account is predicted by my unified analysis of the final rise provides additional support to both.

3.2 The Quantity reading

The Maxim of Quantity requires that the speaker gives the most informative answer that she thinks is true. Therefore, if a speaker indicates, by means of a final rise, that she is violating the Maxim of Quantity, this implies that she, with her final-rising utterance, does not give the most informative answer that she thinks is true, i.e., *that she knows more than she says*. I assume that this lies at the core of the Quantity reading, as typically used in (conjunctive) lists, illustrated by (2).[7,8] At each pre-final list item, the speaker indicates by means of the final rise that she knows more than she has told us so far.[9] Because a violation of Quantity is not very dramatic, the present account predicts that this reading typically occurs with a low final pitch, as seen in the literature (cf. section 2.1).

Just like the Quality reading, the Quantity reading may license additional inferences. Saying less than you know, i.e., violating Quantity, must have a reason. A reason could be that the conversation is between a teacher and a student, where the teacher is not saying everything she knows. For the Quantity violations in a list, a typical reason may be that the speaker is breaking up what she knows into several pieces, giving one at a time, to facilitate reader comprehension: one violates Quantity because it *clashes* with Manner.[10] Alternatively, if no comprehension facilitation is necessary (for instance if the list of people has already been given before), the list in (2a) could be pronounced in a more manner-of-factly way, in a single, falling contour, without any rises:

(5) A: Who came to the party?
B: Mary [high], Bob [mid] and Sue [low]↘.

Existing work on the Quantity reading

To my awareness, regarding the Quantity reading, nothing has been published that goes much beyond the idea that the final rise in lists indicates 'unfinishedness' (e.g., Bolinger, 1982, reiterated in Bartels, 1999; Gunlogson, 2008); I briefly return to this characterisation in section 4). My result suggests how this can be made more precise: list intonation conveys that the speaker knows more (regarding the QUD) than what she has said.

---

[6] Trinh and Crnič say that Truckenbrodt cannot explain why rising declaratives elicit a response while falling declaratives don't. However, as Trinh and Crnič themselves suggest, 'Truckenbrodt could claim that the ability of rising declaratives to elicit a response follows from the speaker not believing that φ and her expressed desire that φ be made common ground: this desire would not be satisfied if the addressee does not utter φ' (p.8). I do not see what they think would be wrong with this suggestion, and I believe nothing is.

[7] I thank in particular Alysson Ettinger and Joseph Tyler for extensive discussion on list intonation.

[8] An anonymous reviewer suggested that lists may have a particular syntactic form with its own intonational norms. However, one would then have to explain why the treatment of the final rise I advocate seems to apply to lists just as well. This need not be hard, but in my view it is unnecessary.

[9] Note that the Quantity reading cannot be what underlies *disjunctive* lists, if such creatures exist at all:

(I) I saw Mary↗. I saw Bob↗. Or I saw Sue↘.

After all, in disjunctive lists, each additional disjunct would *decrease*, rather than increase, the information provided by the speaker. However, the status of the utterance in (I) is unclear to me. It seems somewhat natural only with hesitation markers in between and a puzzled look on the speaker's face, but even then the late occurrence of 'or' rather than 'and' feels slightly surprising. I trust that this can be independently explained in terms of the rhetorical structure of a discourse: the default discourse relation between two subsequent sentences seems to be conjunctive (cf. work on dynamic semantics, in particular SDRT). For this reason, perhaps, a more natural way to express (I) is in a single breath, with a falling contour (no intermediate rises):

(II) I saw Mary [high], Bob [mid], or Sue [low]↘.

In any case, my account of the final rise would predict that a disjunctive list, if a valid discourse strategy at all, requires higher rises than a conjunctive list, indicating (at least) violations of Quality, rather than Quantity. Whether this prediction is borne out is left to future research.

[10] I should emphasize that this implicature, that the speaker is *facilitating* comprehension, does not yet explain why the final rise can be used also for *checking* comprehension, as discussed below in section 3.4.

3.3 The Relation reading

The Maxim of Relation, recall, requires that the speaker's utterance, relative to her information state, entails the QUD. What exactly this implies depends on the meanings assigned to the utterance and the QUD, and the notion of entailment used. Hence, the success of my approach depends in this respect on the *semantics* we assume: it must be such that a violation of the Maxim of Relation yields exactly what I called the 'Relation reading'.

For inspiration, let us consider a pragmatic phenomenon that is intimately connected to what I called the Relation reading: *exhaustivity implicatures*, exemplified in (6).

(6) A: Of John, Mary and Bob, who came?
B: John was there↘. ⇝ *not Mary, not Bob.*

Note that this example is, aside from the final contour, identical to example (3b) of the Relation reading. Where (6), with a final fall, implicates that Mary and Bob weren't at the party, (3b), with a final rise, implies uncertainty about precisely that. (This close connection between exhaustivity and the Relation reading is observed also by Constant (2012), who contrasts rise-fall-rise with 'only'.) This suggests that the Maxim of Relation will be suitable for an account of the Relation reading *if and only if the maxim is strict enough to derive exhaustivity implicatures.* I therefore build on my own recent work on exhaustivity, (Westera, 2013), that derives exhaustivity implicatures via the Maxim of Relation, as discussed next. Afterwards, I show that this indeed accounts for the Relation reading of the final rise.

Westera's (2013) Maxim of Relation

Westera (2013) argues that, for an account of exhaustivity implicatures that solves the problematic 'epistemic step' (Sauerland, 2004) in a wholly Gricean way, the Maxim of Relation must be sensitive to the possibilities that an utterance *draws attention to.* Intuitively, the question in (6) draws attention to the possibility that John came, the possibility that Mary came, and the possibility that Bob came (as well as combinations of these). The response, however, draws attention only to the possibility that John came; it leaves the other possibilities *unattended*, and it is in that sense *not entirely related* to the question. Westera shows that if the Maxim of Relation is sensitive to this, exhaustivity implicatures can be accounted for.

To turn this idea into a formal theory, Westera employs Roelofsen's (2011) *attentive semantics*, which builds on Ciardelli's (2009) *possibility semantics* and subsequent work, in which the meaning of a sentence, called a *proposition*, is a *set of sets of worlds*, i.e., a set of classical propositions. The proposition $[\varphi]$ expressed by a sentence φ is conceived of as the set of possibilities that the sentence *draws attention to.* The union of these possibilities corresponds to the sentence's *informative content*, i.e., the information provided by the sentence, which is treated wholly classically. I adopt the following notions and notations:

- *Informative content*: $|\varphi| := \bigcup[\varphi]$
- *A restricted to* a set of worlds s:
  $A_s := \{\alpha \cap s \mid \alpha \in A, \alpha \cap s \neq \varnothing\}$

For the relevant fragment of propositional logic, the semantics is defined recursively as follows:

1. $[p] \quad = \{\{w \in \textbf{Worlds} \mid w(p) = \mathsf{true}\}\}$
2. $[\neg\varphi] \quad = \{\overline{\bigcup[\varphi]} \mid \overline{\bigcup[\varphi]} \neq \varnothing\}$
3. $[\varphi \vee \psi] = ([\varphi] \cup [\psi])_{|\varphi| \cup |\psi|}$ $(= [\varphi] \cup [\psi])$
4. $[\varphi \wedge \psi] = ([\varphi] \cup [\psi])_{|\varphi| \cap |\psi|}$

With this richer-than-usual semantics, entailment becomes sparser than usual:

(7) *A entails Q*, $A \vDash Q$, iff:
a. $\bigcup A \subseteq \bigcup Q$; and
b. $Q_{\bigcup A} \subseteq A$.

Item a. requires, just like classical entailment, that A is *at least as informative* as Q. Item b. requires that A is, in addition, *at least as attentive* as Q. That means that every possibility that Q draws attention to, must be a possibility that A draws attention to, insofar as this is compatible with the information provided by A.

This notion of entailment is plugged into the Maxim of Relation, as assumed in section 2.2:

(8) For a cooperative speaker with information s, responding A to Q:
Relation: $A_s \vDash Q$.

From (7) it follows that this maxim requires that every possibility in Q that is not in A, i.e., every possibility that A *leaves unattended*, must, given the speaker's information s, either be incompatible with A, or coincide with a possibility in A.[11]

---

[11] Westera (2013) gives equally formal implementations of the maxims of Quality and Quantity, based on attentive semantics, which would have derived exactly the Quality and Quantity readings discussed above. For the present purposes, however, such formal rigour was unnecessary, because for the maxims of Quality and Quantity, the step from intuition to formalisation is much more direct.

Deriving the Relation reading

Now, example (3a) is accounted for as follows. Let the question (whether John was at the party) translate as $p \vee \neg p$, and the response (that it was raining) as r. These have the following meanings:

(9) $[p \vee \neg p] = \{|p|, \overline{|p|}\}$; $\quad [r] = \{|r|\}$

For the response to be related to the question, both $|p|$ and $|\neg p|$ must, relative to the speaker's information and the information that r, either coincide with $|r|$ or be incompatible with it, i.e., be included in $|\neg r|$. This requirement can be met in two ways:

- The speaker thinks that if it rained, John was there ($s \subseteq \overline{|r|} \cup |p|$; the response restricted to this information yields $\{|r| \cap |p|\}$, which entails $\{|p|, \overline{|p|}\}$); or
- The speaker thinks that if it rained, John wasn't there ($s \subseteq \overline{|r|} \cup \overline{|p|}$; the response restricted to this information yields $\{|r| \cap \overline{|p|}\}$, which entails $\{|p|, \overline{|p|}\}$).

If the final rise conveys a violation of the Maxim of Relation, that means neither of these requirements can be met, i.e., that the speaker does not know how John's attendance depended on the rain ($s \nsubseteq \overline{|r|} \cup |p|$ and $s \nsubseteq \overline{|r|} \cup \overline{|p|}$). This is the Relation reading for example (3a). Recall from section 2 that, despite the maxim violation, the speaker is still presumed to be cooperative. That explains why (3a) is odd unless the responder suspects that the hearer may know of a dependency between the weather and John's attendance.[12]

Example (3b) is also accounted for. As in (Westera, 2013), I assume that the question, for each combination of individuals, draws attention to the possibility that they came, as well as the possibility that no one came. For simplicity, and without loss of generality, I consider only the possibilities that John came, that Mary came, that Bob came, and that no one came. Let p, q and r translate that John, Mary and Bob came, respectively. The question and response then become:

(10) $[p \vee q \vee r \vee (\neg p \wedge \neg q \wedge \neg r)]$
$= \{|p|, |q|, |r|, \overline{|p|} \cap \overline{|q|} \cap \overline{|r|}\}$; $\quad [p] = \{|p|\}$

For the response to be related to the question, each of the question's possibilities must, relative to the speaker's information and the information that p, either coincide with $|p|$ or be incompatible with it. For $|p|$ as well as $\overline{|p|} \cap \overline{|q|} \cap \overline{|r|}$, this is already complied with. For Mary ($|q|$), however, which is 'left unattended' by the response, this means that:

- The speaker thinks that if John was there, Mary was there ($s \subseteq \overline{|p|} \cup |q|$); or
- The speaker thinks that if John was there, Mary wasn't there ($s \subseteq \overline{|p|} \cup \overline{|q|}$)

And likewise for Bob ($|r|$). If for each of Mary and Bob, one of these requirements would be met, then the response, together with this information, would entail the question, i.e., it would comply with the Maxim of Relation. In 'normal' circumstances, i.e., where no maxim is violated, these requirements would enable one to take the *epistemic step* and derive exhaustivity implicatures (Westera, 2013). In the present case, if the final rise conveys a *violation* of the Maxim of Relation, this means that for either Mary or Bob, and possibly both, neither of these requirements can be met. This implies that for Mary or Bob, the speaker does not know how their presence depended on John's presence. Since the speaker thinks John was present ($s \subseteq |p|$), she must not know whether Mary came or she must not know whether Bob came ($s \nsubseteq |q|$ and $s \nsubseteq \overline{|q|}$, or $s \nsubseteq |r|$ and $s \nsubseteq \overline{|r|}$). This is the Relation reading for example (3b).

Summing up: if an utterance leaves one of the QUD's possibilities unattended, the Maxim of Relation requires that the speaker knows how it depends on the information that the speaker did provide. A *violation* of the Maxim of Relation thus entails that there is at least one possibility in the question, of which the speaker does *not* know how it depends on the information she provided. Together with the usual Quality implicature this yields the Relation reading: that there is some possibility in the QUD about which the speaker is uncertain. Finally, because a violation of the Maxim of Relation isn't grave, it is predicted that the Relation reading typically occurs with a low rise, as observed in the literature (section 2.1).

Existing work on the Relation reading

I will compare my account of the Relation reading to two theories of rise-fall-rise, old and new, namely Ward and Hirschberg's (1985) (which is very close to the present approach) and Constant's (2012) (which criticizes the former).

---

[12] I thank an anonymous reviewer for pointing out that this had to be made more explicit. She says that in (3a), 'B must be uncertain about whether it really has something to do with the question, but must suspect that for A it might have.'

Ward and Hirschberg (1985) For Ward and Hirschberg, rise-fall-rise intuitively conveys uncertain relevance. They make this more precise by assuming that rise-fall-rise conveys one of three types of 'scalar uncertainty', about (i) whether it is appropriate to evoke a scale at all, (ii) which scale to choose, given that some scale is appropriate, and (iii) given some scale, uncertainty about the choice of some value on that scale. Now, we find their distinction between type (ii) and type (iii) rather moot; both are illustrated with examples like (3a), except that for type (iii) the examples they use are more evidently scalar.[13] For instance, while they use an example like (3a) for type (ii), they present (11) as an example of type (iii) (adapted from their (60)):

(11) A: Does your friend live far away?
B: In suburban Philadelphia ↗.

Here, B is unsure whether suburban Philadephia corresponds to 'far' on the distance scale. However, Ward and Hirschberg use the word 'scale' rather liberally, meaning, roughly, 'QUD'. Hence, it is easy to frame example (3a) in exactly the same way: B is unsure whether that it was raining corresponds to 'yes' on John's attendance scale. Hence, I believe that their type (ii) and type (iii) readings can be conflated. Indeed, my account derives the Relation reading for both examples alike.

As for their type (i) uncertainty, I think this is genuinely a different reading. They illustrate it with the following example (their (52)):

(12) A: Do you speak Ladino?
B: I speak Spanish ↗.

As Ward and Hirschberg explain, here B conveys uncertainty about whether A is interested only in Ladino, or whether other Iberian languages are also relevant. Keeping in mind Ward and Hirschberg's liberal use of 'scale' as meaning 'QUD', I understand this as a case in which B conveys that she is uncertain about *what the QUD is*. Now, in my derivation of the Relation reading, I have, so far, implicitly assumed that the speaker knows what the QUD is. But of course, one way of failing to know how one's utterance relates to the current QUD, is to not know what the current QUD is to begin with. Hence, the present account already predicts that the final rise, if it conveys a violation of the Maxim of Relation, can convey this kind of uncertainty, too.[14]

Constant (2012) Constant assumes that rise-fall-rise is a 'universal quantifier of assertable alternative unclaimability' (p.39). That is, rise-fall-rise on a sentence φ universally quantifies over φ's alternatives (say, answers to the QUD) that are neither entailed nor excluded by φ itself, of which there must be at least one, and says of these that the speaker lacks the information to support them. Before evaluating this approach, it is worth noting, as Constant himself does, that it solves most puzzles he discusses purely due to the requirement that the quantification is non-vacuous, i.e., that there is at least one non-excluded, non-entailed alternative. Since my account predicts that the Relation reading has existential force (e.g., in (3b), the speaker is unsure about *someone* of Mary and Bob), it inherits from Constant those solutions.

Crucially, Constant assumes that rise-fall-rise signifies not uncertainty but, merely, a lack of belief that the alternatives are true. This would mean that rise-fall-rise would be compatible with the speaker believing that all alternatives are false, i.e., with an exhaustivity implicature - which it isn't.[15] Indeed, the contribution of rise-fall-rise according to Constant would be *equivalent to the standard Quantity implicature*. Since exhaustivity as a conversational implicature is derived through the Maxim of Quantity, promoting the Quantity implicature to a semantic entailment should, if anything, make the exhaustivity implicature *more* salient. For this reason, I believe that Constant's account of the final rise is too weak.

Nevertheless, let us consider the example used by Constant to motivate this weakness (his (60), adapted from Oshima, 2005):

---

[13] In addition, Ward and Hirschberg classify certain examples as type (ii) that Constant (2012) argues are in fact 'metalinguistic'. I will not discuss those at present.

[14] Another example of this kind of Relation violation is the following, where, as pointed out to me by an anonymous reviewer, the final rise is taken to contribute a query as to what exactly the receptionist's question is, i.e., along which properties the question should be taken to divide the logical space:

(III) (Customer approaches hotel receptionist)
Receptionist: Who are you?
Customer: I'm John Smith ↗.

[15] If one assumes, instead, that exhaustivity is not a conversational implicature at all, but, rather, due to a 'silent only' operator (e.g. Chierchia, Fox, & Spector, 2012), this objection would not necessarily hold, because rise-fall-rise, for Constant, requires that there are non-dispelled alternatives - and *grammatical* exhaustivity would dispel them all. I will not explore this option, for reasons discussed by Westera (2013).

(13) A: Did your friends pass the test?
B: John passed↗. Bob and Sue flunked.

Here, Constant says, rise-fall-rise occurs despite B not being uncertain about Bob and Sue, and this would be problematic for Ward and Hirschberg (1985). But in defence of Ward and Hirschberg (and myself), I object that the alternatives to which the final rise pertains here are not Bob and Sue, but, rather, whether B's friends passed the test or not (i.e., the answers to the QUD). Now, it is known of plural indefinites that, when some-but-not-all of B's friends passed, the sentence 'B's friends passed' is judged neither true nor false (e.g. Landman, 1989). Hence B, in uttering that John passed while knowing that Bob and Sue flunked, can be genuinely uncertain as to whether this corresponds to a 'yes'-answer or a 'no'-answer. This is what licenses the rise in (13).[16]

In sum, I think that Constant's (2012) account is too weak, and that the example he uses in favour of this weakness may have a different explanation.

3.4 A 'Manner' reading

So far I have discussed *three* readings, whereas I distinguished *four* maxims. This suggests that a fourth reading, a 'Manner reading', should exist. The Maxim of Manner requires that the speaker thinks she is making herself understood, hence its *violation* would imply that the speaker *lacks this belief.* This suggests that the final rise can be used for *comprehension checking*, a use which indeed surfaces in the literature, linked to features such as politeness (e.g., Gussenhoven, 2004).

However, the Manner reading is difficult to isolate. For instance, does mispronouncing a name or technical term make a statement false (Quality), or unclear (Manner)? And if one is uncertain about the particular wording of one's answer, is this uncertainty about Manner, or about what exactly the QUD is (Relation)? Despite this blurriness, I think the following example may succeed at isolating a pure Manner reading:

(14) (English tourist in a French café.)
I'd like... err... je veux... black coffee↗.

Given that the tourist knows what she wants, and that it is available, the final rise cannot convey uncertainty about the proposition expressed (Quality). She also cannot be uncertain about what question she is addressing (Relation). If we assume that black coffee is all she wants (and that she is alone), a Quantity (list) reading is also ruled out, and the rise can really only pertain to her uncertainty as to whether she made herself understood: Manner. If she considers it likely that she was understood, a low rise is predicted. However, Manner violations could in principle be as dramatic as Quality violations, given that making oneself understood is a precondition for conveying any kind of content at all.

4 Predictive power

I wish to discuss, and hope to dispel, three worries regarding the predictive power of my proposal. First, one might wonder whether my theory is not *too general.* Since the set of maxims is in principle open-ended, it may seem that there are practically no constraints on what a final rise may be used to convey. However, this lack of constraints is only apparent. Any maxim must be thoroughly motivated as a general principle of rational communication. Hence, while my theory does not constrain the *number* of different readings a final rise may have, it does very rigidly constrain the *kinds* of readings that it may have: any reading should be understandable in terms of the violation of some rule of rational communication. This enforces a particular *mode of explanation* for any new use of the final rise that might be discovered (just like Grice's theory of pragmatics invites a particular mode of explanation for implicated content). My theory would be falsified (or its generality challenged) if some use of the final rise is found that cannot be understood as the violation of a maxim (or, conversely, if some maxim violation is discovered that cannot be conveyed by a final rise).

Second, one might wonder whether my account can predict, for a given utterance, which of the many uses of the final rise is intended. The answer is 'no, not on its own'. However, it does make very specific predictions as to what each of the readings exactly pertains to. If we add to these predictions a bit of contextual knowledge, then the ambiguity is easily resolved. Consider the following example (suggested to me by an anonymous reviewer; it is similar to many examples discussed by Ward and

---

[16] To test this explanation, one may compare (13) to (IV):

(IV) A: Did *all* of your friends pass the test?
B: ?? John passed↗. Bob and Sue flunked.

It seems to me that the final rise is strange here (if we rule out a Quality or Quantity reading), because the speaker *does* know how to answer the question: with a clear 'no'.

Hirschberg (1985); Constant (2012)):

(15) A: Are you rich?
B: I'm a millionaire↗. (low rise)

The following readings are predicted:

- **Quality:** B is uncertain about her being a millionaire, and either depressed or very uninterested in finding out the truth of this proposition (because the rise is low).
- **Quantity:** B knows more about A's question than she said. Perhaps B is giving A a very obvious hint; or perhaps B happens to be a *billion*are reluctant to reveal it.
- **Relation:** B is unsure about how this resolves A's question. This can only be if B is unsure about A's intended interpretation of 'rich' (say, because A is a *billion*aire).
- **Manner:** B is unsure how to pronounce 'millionaire', or whether A knows the word.

Now, each of these readings is indeed *possible*, which shows that a theory as general as the present one is really necessary. But let us now add some plausible assumptions about the context. If A and B are both native speakers of English, the Manner reading is ruled out. If, in addition, B knows approximately how rich she is, which is likely, then the Quality reading is ruled out. Furthermore, if neither A nor B is a *billion*aire, then the only reading that makes sense is the Quantity reading, explained as B giving A a very obvious hint, perhaps because B is slightly annoyed by A's stupid question. In sum, the ambiguity is quite easily resolved by contextual knowledge. Should context prove insufficient, then various linguistic (including gestural) tools may aid in disambiguating the final rise, as mentioned in footnote 2 (section 2).[17]

The third and final worry I wish to discuss is whether the theory outlined here is even *general enough*. Since my theory leaves a lot of disambiguating to be done anyway, why not say that the final rise conveys a general, underspecified 'unfinishedness', as proposed for instance by Bolinger (1982), and let other intonational, contextual, or gestural features fill in the blanks? The reason is that, for this alternative theory to yield any testable predictions, one would have to specify in what sense or senses an utterance might be 'unfinished'. I am confident that, if one attemps this in an empirically accurate way, one will end up defining 'unfinishedness' as something like 'by itself not a cooperative contribution to the discourse' - and this is not at all different from what I have proposed.

5 Conclusion and outlook

I have analysed the final rise on declaratives in English as indicating that a maxim is being violated; i.e., it negates exactly that which, according to Grice, is supposed in conversation. This analysis is *unifying*, in the sense that (i) it captures intuitions found in existing work, (ii) it relies on machinery (e.g., the conversational maxims) that comes straight from the literature; and (iii) existing but thus far disconnected accounts of different uses of the final rise were reproduced, predicting four salient readings: Quality, Quantity, Relation and Manner. Crucial for the Relation reading was the Maxim of Relation's sensitivity to attentive content, motivated by the link between the Relation reading and exhaustivity implicatures.

Given the importance of marking the violation of a maxim (so as not to mislead), the function carried in English by the final rise is expected to be realized cross-linguistically, whether by intonation, discourse particles (especially in *tonal* languages where, as an anonymous reviewer remarks, the intonation channel is unavailable), or other means. I suspect that the same method of using pragmatic notions within a semantic specification is also applicable there. It will be interesting to see to what extent, cross-linguistically, the four readings are expressed by a single construction, as in English, or whether they are subdivided in particular ways. This would provide a window on whether the four Gricean maxims reflect in any way how language users decompose the notion of cooperativity.

In the future I hope to extend the present theory to the domain of rising and falling *interrogatives*. But first, current work in progress is aimed at extending the theory to the notion of *contrastive topic* (Büring, 2003). Contrastive topic, associated with a pitch accent in a rising intonation phrase, is generally thought to indicate that the speaker targets only a subquestion of some overarching QUD. This can be analysed as a violation of Quantity or Relation regarding the overarching QUD, while, as far as the subquestion is concerned, the speaker may fully comply with the maxims.

---

[17] There is quite a salient, humorous 'pretense' Relation reading for (15): B would be jokingly pretending not to know whether millionaires are considered rich. Probably, what disambiguates between this humorous reading and the 'obvious hint' Quantity reading, is a wink or a smirk.

Acknowledgements

I am very grateful to J. Groenendijk, I. Ciardelli, F. Roelofsen, J. Shaheen, J. Tyler, A. Ettinger, M. Križ, the audiences of S-Circle (Santa Cruz, 04/2013) and CISI (Düsseldorf, 08/2013), and the anonymous reviewers for SemDial (Amsterdam, 12/2013) for valuable comments. Financial support from the Netherlands Organisation for Scientific Research (NWO) is gratefully acknowledged.

References

Banziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*, 252-267.

Bartels, C. (1999). *The intonation of English statements and questions: a compositional interpretation*. Routledge.

Bolinger, D. (1982). Intonation and its parts. *Language*, *58*, 505-533.

Büring, D. (2003). On d-trees, beans, and accents. *Linguistics and Philosophy*, *26*, 511-545.

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In C. Maienborn, P. Portner, & K. von Heusinger (Eds.), *Semantics: An international handbook of natural language meaning* (Vol. 2, p. 22972332). Mouton de Gruyter.

Ciardelli, I. (2009). *Inquisitive semantics and intermediate logics.* (Master Thesis, University of Amsterdam)

Constant, N. (2012). English rise-fall-rise: a study in the semantics and pragmatics of intonation. *Linguistics and Philosophy*, *35(5)*, 407-442.

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58).

Groenendijk, J., & Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers*. Unpublished doctoral dissertation, University of Amsterdam.

Gunlogson, C. (2003). *True to form: Rising and falling declaratives as questions in english*. Routledge.

Gunlogson, C. (2008). A question of commitment. *Belgian Journal of Linguistics*, *22*, 101-136.

Gussenhoven, C. (1983). *A semantic analysis of the nuclear tones of english*. Indiana University Linguistics Club.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Landman, F. (1989). Groups I and II. *Linguistics and Philosophy*, *12*, 559–606, 723–744.

Oshima, D. Y. (2005). Morphological vs. phonological contrastive topic marking. *Chicago Linguistic Society*, *41*.

Roberts, C. (1996). Information structure in discourse. In J. Yoon & A. Kathol (Eds.), *Osu working papers in linguistics* (Vol. 49, pp. 91–136). Ohio State University.

Roelofsen, F. (2011). *Information and attention.* (Manuscript, ILLC University of Amsterdam)

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*(3), 367–391.

Schulz, K., & Van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, *29*, 205–250.

Trinh, T., & Crnič, L. (2011). On the rise and fall of declaratives. In I. . Reich (Ed.), *Proceedings of the conference "SuB 15 – Sinn und Bedeutung"* (p. 1-16). Universaar - Saarland University Press.

Truckenbrodt, H. (2006). On the semantic motivation of syntactic verb movement to C in German. *Theoretical Linguistics*, *32*(3), 257–306.

Tyler, J. (2012). *Discourse prosody in production and perception*. Unpublished doctoral dissertation, University of Michigan.

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: the pragmatics of fall-rise intonation. *Language*, *61.4*, 747-776.

Ward, G., & Hirschberg, J. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in english. *Journal of Phonetics*, *20*, 241-251.

Westera, M. (2013). Exhaustivity through the maxim of Relation. In *Proceedings of LENLS10* (p. 193-205).

Multimodal Prediction of Psychological Disorders: Learning Verbal and Nonverbal Commonalities in Adjacency Pairs

Zhou Yu* **Stefen Scherer**† **David Devault**† **Jonathan Gratch**†

Giota Stratou† **Louis-Philippe Morency**† **Justine Cassell***

*School of computer Science
Carnegie Mellon University
{zhouyu,justine}@cs.cmu.edu

†Institute for Creative Technology
University of Southern California
{scherer,devault,gratch,
stratou,morency}@ict.usc.edu

Abstract

Semi-structured interviews are widely used in medical settings to gather information from individuals about psychological disorders, such as depression or anxiety. These interviews typically consist of a series of question and response pairs, which we refer to as *adjacency pairs*. We propose a computational model, the Multimodal HCRF, that considers the commonalities among adjacency pairs and information from multiple modalities to infer the psychological states of the interviewees. We collect data and perform experiments on a human to virtual human interaction data set. Our multimodal approach gives a significant advantage over conventional holistic approaches which ignore the adjacency pair context in predicting depression from semi-structured interviews.

1 Introduction

Recent advances in the fields of verbal and nonverbal behavior analysis are revolutionizing our ability to analyze and understand people's behavior. One promising application is the automatic analysis of nonverbal behaviors associated with psychological disorder. Extensive research in behavioral sciences has demonstrated a link between specific psychological disorders, such as depression, and patterns of verbal and nonverbal behavior (Ellgring, 1989). Recognizing these verbal and nonverbal indicators, however, requires expert judgements from trained clinicians. The factors underlying these judgements are not easily quantifiable (Ellgring, 1989). Automatic detection of verbal and nonverbal indicators can assist clinicians by supporting their interview processes and providing more systematic, quantified measurements. Moreover, fully-automated techniques can serve as a pre-screening instrument for patients, complementing the self-reported questionnaires which are currently used for this purpose.

Psychological assessment interviews consist of a series of "question" and "response" pairs, which are consecutive utterances that we refer to as adjacency pairs (Schegloff, 2007). The potential "response" doesn't have to be a direct answer, but could be a counter-question or other form of response triggered by the "question", as long as it satisfies Grice's conversational maxim of relevance (Grice, 1975). Different adjacency pairs serve different purposes in triggering subject responses, and a model that considers context could better predict psychological disorders. We propose a computational approach that leverages the advantage of verbal and nonverbal behaviors extracted at the adjacency pair level to support a more contextualized analysis, unlike previous approaches which ignored context (Cohn et al., 2009), or only consider context in single feature analysis (DeVault et al., 2013).

Based on Hidden Conditional Random Fields (HCRFs) (Quattoni et al., 2004), we propose a new computational model, the Multimodal HCRF. HCRFs allow us to learn verbal and nonverbal commonalities among adjacency pairs automatically. For example, one specific commonality is that depressed people have a lower speech rate compared to non-depressed people in their responses to a large set of probing questions (see section 8.4 for details). In order to assess the effectiveness of incorporating adjacency pair into our analysis, we performed experiments on a corpus of 130 human to virtual human interviews, where the question was always asked by the virtual human interviewer, and the response was given by the real human. Our analysis relies on a model which brings together behaviors from multiple modalities: visual, acoustic and conversational and results showed a significant improvement for our multimodal computational model over previous models at predicting depression.

We first review previous work and our hypotheses before we describe our dyadic interaction data set. After that we introduce automatically extracted multimodal features that capture verbal and nonverbal behaviors. Next, we present our computational model and experiments to validate it. Finally, we further analyze the results from our experiments.

2 Related Work

Many previous studies have examined the links between nonverbal behaviors and clinical conditions (Ellgring, 1989; Cohn et al., 2009). Little progress has been made towards identifying any clear links between patient disorders and expressed behaviors. This is due to the difficulties of manually annotating gestures and facial expressions, inconsistent measurements of nonverbal behaviors across studies and differences in social contexts of the interview processes between studies.

There is a general consensus regarding the relationship between certain clinical conditions (especially depression and social anxiety) and associated verbal and nonverbal cues. Emotional expressivity, such as the frequency or duration of smiles, is diagnostic of psychological disorders. For example, depressed patients frequently display flattened or negative effects, including less emotional expressivity (Perez and Riggio, 2003; Bylsma et al., 2008), fewer mouth movements (Fairbanks et al., 1982; Schelde, 1998), more frowns (Fairbanks et al., 1982; Perez and Riggio, 2003) and fewer gestures (Hall et al., 1995; Perez and Riggio, 2003). Some findings suggest that the quantity of expressions may not be as important as their dynamics. For example, depressed patients may frequently smile, but these smiles are perceived as less genuine and often shorter in duration (Kirsch and Brunnhuber, 2007). Social anxiety and PTSD share some features with depression, such as a tendency for heightened emotional sensitivity and more energetic responses. Such responses can include startlement and a greater tendency to display anger (Kirsch and Brunnhuber, 2007) or shame (Menke, 2011). Cohn and colleagues have identified increased speaker-switch durations as indicators of depression, and have explored the use of these features for classification (Cohn et al., 2009). Our current research builds on these findings as a step to overcome the difficulty of manually annotating human behavior.

Scherer et al. (2013b) explore the correlation between automatically quantified acoustic and visual features with psychological disorders. Stratou et al. (2013) find that the subject's gender plays an important role in automatic assessment of psychological conditions when analyzing automatically extracted visual features. DeVault et al. (2013) investigate the correlation between conversation features and psychological disorders, but don't take visual features into consideration. Cohn et al. (2009) use both facial expression and vocal prosody in identifying depression, however, they do not include more features which are predictive of depression. In summary, there is a lack of models that combine comprehensive conversational, visual and acoustic features related to depression. Also, the prediction methods used in previous works do not take the contextual information of the interview into account.

We include contextual information by modeling nonverbal behavior at the adjacency pair level. We apply HCRFs for classification, as opposed to Naive Bayes used in DeVault et al. (2013) and Stratou et al. (2013) because HCRFs model time contingency. HCRFs have been successfully used to tackled problems in computational vision and speech. For instance, Quattoni et al. (2004) applied HCRFs to model spatial dependencies for object recognition in unsegmented cluttered images.

3 Research Hypotheses

Interviews typically consist of a series of question and response pairs which we refer to as *adjacency pairs*. We present the two consecutive utterances as a tuple (q_i, r_i), where q is the "question" and r is the "response".

For each adjacency pair, subjects exhibit different verbal and nonverbal behaviors, for example, a different speech rate or facial expression. We hypothesize that:

1. We can better predict depression with a computational model that takes advantage of context by considering features quantified at the **adjacency pair level** rather than models using features extracted from the whole interaction. For example, we consider the speech rate in the response of the subjects in different adjacency pairs as opposed to the speech rate over the whole interaction in our model. The change of nonverbal behaviors exhibited in human responses to different stimuli (i.e. positive questions versus negative questions) are known to be significantly different between groups with and without psychological

disorders (Bylsma et al., 2008).

2. Adjacency pairs which serve the same probing purpose **share commonalities** in human verbal and nonverbal responses. By allowing our model to learn these commonalities we can improve prediction accuracy. For example, one commonality could be that for a set of adjacency pairs which concern a client's personal experience, people with psychological disorders have a longer latency in speech onset time to respond to the questions.

3. A comprehensive set of features from **multiple modalities** improves computational performance in predicting depression compared to a single or bi-modal approach. Previous works (Cohn et al., 2009; Scherer et al., 2013b; Stratou et al., 2013) combine different multimodal features, but none of these approaches make use of all three modalities (conversational, visual and acoustic). According to our previous research, multimodal features also improve friendship prediction (Yu et al., 2013). Although the tasks are different, we believe that leveraging multiple information channels can benefit depression prediction.

4 Distress Assessment Interview Corpus (DAIC)

We use a data set that has 130 semi-structured interviews in a Wizard-of-Oz paradigm between a human and the virtual character Ellie, depicted in Figure 1. Drawing on observations of interviewer behavior in the face-to-face dialogues, Ellie was designed to serve as an interviewer who is also a good listener, providing empathetic responses, back channels and continuation prompts to elicit extended replies to specific questions. The virtual human builds rapport with the participant at the beginning of the interaction with a series of casual questions about Los Angeles. After that, the conversation transitions towards intimate questions, like, "*Do you consider yourself more shy or outgoing?*". After the intimate phase, Ellie asks questions directly related to previous experiences of psychological disorders, such as, "*Have you been diagnosed with depression before*?". A series of positive questions, for example, "*How would your best friend describe you?*" are designed to leave the participant in a positive mood. Participants for the study were recruited via Craigslist and all applicants who met the requirements (i.e. age greater than 18, and adequate eyesight) were accepted. The mean age of the 130 participants in our data set was 38.41 years, with 69 males and 61 females. For a measure of psychological disorders, the PHQ-9 provides guidelines on how to assess the participants' conditions based on their responses to a questionnaire. Among the 130 participants, according to the PHQ-9, 30 participants were considered to have moderate depression or above (Kroenke and Spitzer, 2002) by having a cumulative score of ten or above. We consider them depression-positive in this study.

Figure 1: Ellie, the virtual human

5 Automatically Extracted Multimodal Features

In this section, we briefly describe the features used in our experiments and the literature that motivates them. We focus on three types of features: conversational (Section 5.1), visual (Section 5.2) and acoustic (Section 5.3). All the features are extracted from the "response" part of an adjacency pair, as the "question" part of an adjacency pair is spoken by Ellie and is identical for all the subjects. We include only automatically derivable features in our analysis for the purpose of reducing manual annotation. In total, we use 16 features: 5 conversational, 3 visual and 8 acoustic.

5.1 Conversational Features

The system's speech segments, including starting and ending time stamps and verbatim transcripts of system utterances, were saved from the system log files. Motivated by DeVault et al. (2013), we selected the following features:

- **Speaking Rate and Onset Time** Slowed speech and increased onset time were observed in previous clinical interviews of depressed individuals (Hall et al., 1995). We quantify the speaking rate by counting the number of words spoken per minute, and the speech onset time as the time delay before the user responds to Ellie's question. Here we use the manual transcription of the interview. However, it is possible for the output of the automatic speech recognition (ASR) system to be used as an approximation of the transcription, thus making the speech rate and onset time automatically obtainable.

- **Number and Average Length of User Segments** The utterances are automatically segmented by identifying long pauses and the average length of the user segments is quantified in seconds.

- **Filled Pause Rate** We count the number of times any of the tokens uh, um, uhh, umm, mm, or mmm appears in each speech segment. To account for the varying length of speech segments, we define the filled pause rate as the number of those tokens divided by the duration of the corresponding segment.

5.2 Visual Features

We selected three visual features based on work in Stratou et al. (2013):

- **Expression Variability** Based on a collection of clinical observations summarized in Ellgring (1989), the homogeneity of an affective level and total facial activity are considered good indicators of psychological disorders. Specifically, reduced facial behavior, or lack of emotional variability, has been reported as an indicator of depression. Our automatic feature extraction system includes the Computer Expression Recognition Toolbox (CERT) (Littlewort et al., 2011), which measures 8 basic expressions: Anger, Disgust, Contempt, Fear, Joy, Surprise, Sadness and Neutral. We measure emotional variability by considering the variances of all these expressions.

- **Neutral Expression** The frequency of the detection by CERT of a "Neutral" expression is a good measure of emotional "flatness", which mentioned in Ellgring (1989) as well.

- **Head Rotation** Clinical observations suggest reduced motor variability or motor retardation among patients suffering from depression (Ellgring, 1989). Hence, as an aspect of motor variability we look at head rotation variability as an indicator of psychological disorders. Our system for automatic analysis provides 3D head position and orientation based on the GAVAM head tracker (Morency et al., 2008) and CLM-Z face tracker (Baltrusaitis et al., 2012). Measuring the head rotation in all three directions (yaw, tilt and roll) allows us to calculate the head rotation.

5.3 Acoustic Features

Motivated by Scherer et al. (2013a) and Cohn et al. (2009), we extracted the following acoustic features with a sample rate of 100 Hz, using the lapel microphone recordings:

- **Energy in dB** The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms (i.e. 100Hz sample rate). Each speech window is filtered with a hamming window and the energy is calculated and converted to the dB-scale.

- **Fundamental Frequency** **(f_0)** In Drugman and Abeer (2011), a method for f_0 tracking based on residual harmonics, which is especially suitable in noisy conditions, is introduced. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering. This process reduces the influence of noise and vocal tract resonances. For each $r(t)$, the amplitude spectrum is computed, showing peaks for the harmonics of f_0, the fundamental frequency. These peaks form the basis for robust f_0 estimation.

- **Spectral Stationarity** **(ss)** To characterize the range of the prosodic inventory used over utterances, we make use of the so-called *spectral stationarity* measure. This measurement was used in Talkin (1995) as a way of modulating the transition cost used in the dynamic programming method used for f_0 tracking. Spectral stationarity, ss, is measured using the Itakura distortion measure (Itakura, 1975) between the current current and previous frame. We use a relatively long frame length of 60 ms (with a shift of 10 ms; sampling rate 100Hz) and frames are windowed with a Hamming window function before measuring ss.

- **Normalized Amplitude Quotient (NAQ)** This feature is derived from the glottal source signal estimated by iterative adaptive inverse filtering (Alku et al., 1992). The output is the differentiated glottal flow. The NAQ is the ratio between the negative amplitude of the main excitation in the differentiated glottal flow pulse and the peak amplitude of the glottal flow pulse normalized by the length of the glottal pulse period (Alku et al., 2002).

- **Quasi-Open Quotient (QOQ) and Open-Quotient Neural Net (OQ$_{NN}$):** The QOQ is also derived from amplitude measurements of the glottal flow pulse (Alku et al., 2002). The quasi-open period is measured by detecting the peak in the glottal flow and finding the time points before and after this point that descend below 50% of the peak amplitude. The duration between these two time-points is divided by the local glottal period to get the QOQ parameter. As a novel alternative of the QOQ, we extract **OQ**$_{NN}$, a parameter estimating the open quotient using standard Mel frequency cepstral coefficients and a trained neural network for open quotient approximation (Kane et al., 2013).

- **Harmonic Amplitude Difference** The difference in amplitude levels (in dB) between the first two harmonics of the narrow band voice source spectrum, which is an alternative rough estimate of the open quotient (Henrich et al., 2001).

- **Peak Slope** This voice quality parameter is based on features derived following a wavelet-based decomposition of the speech signal (Kane and Gobl, 2011). The parameter, named *peak*, is designed to identify glottal closure instances from glottal pulses with different closure characteristics.

6 The Multimodal HCRF Modal

A semi-structured interview changes according to the behaviors of the participants and is composed of a series of adjacency pairs. From a modeling perspective, semi-structured interviews have three main components: (1) an overall goal, which is specific to each interview (e.g., assessing depression or PTSD), (2) a conversational structure where some adjacency pairs share a common purpose and (3) a variation in human behavior during different adjacency pairs or sets of adjacency pairs. We propose a computational approach which explicitly models these three main components and addresses all the research hypotheses discussed in Section 3. Our approach is based on a Hidden Conditional Random Field (HCRF) (Quattoni et al., 2007) which is a probabilistic energy model that learns hidden commonalities automatically from a series of observations from adjacency pairs and their corresponding mappings to depression assessments. Each hidden state groups together adjacency pairs with similar function for the purpose of differentiating depressed people from non-depressed. We propose to adapt HCRF to automatically predict depression over the semi-structured interviews between humans and virtual humans.

Figure 2 depicts a graphical representation of our model. We wish to learn a mapping between multimodal features $\boldsymbol{x}$=$\{x_1, x_2, ..., x_n\}$, defined in Section 4 and extracted at the adjacency pair level, and the class label y $\in$ Y, which is either depressed or not. Our model is defined as

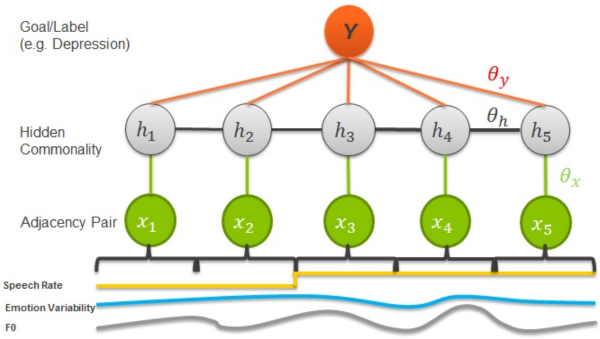


Figure 2: Multimodal HCRF

$$P(y|\boldsymbol{x},\theta) = \frac{\sum_{\mathbf{h}} e^{\psi(y,\mathbf{h},\mathbf{x};\theta)}}{Z(\boldsymbol{x},y)}$$

where **h** = $\{h_1, h_2, ..., h_m\}$ are hidden states representing the commonalities between adjacency pairs. H is the set of hidden commonalities. The constant Z($\boldsymbol{x}$,y) is a partition function that serves as a normalization factor. The most important parts of the model are the potential functions, $\psi(y, \mathbf{h}, \mathbf{x}; \theta)$, parameterized by [θ_x θ_y θ_h]. We visualize these parameters in Figure 2 and describe them below:

1. The parameter θ_x models the relationship between multimodal features x_j and hidden states (commonalities) h_j. By analyzing the

amplitude of each of the weights in θ_x, it is possible to learn the relative importance of each feature for each hidden state. Adjacency pairs that map to the same hidden state form a group which share commonalities.

2. The parameter θ_y models the relationship between the hidden states h_j and the label y. By analyzing the weights of θ_y, it is possible to see which groups of adjacency pairs are important to predict depression.

3. The parameter θ_h represents the links between hidden states. It models the temporal dynamics in the hidden states (commonalities) of adjacency pairs.

In our experiments we used a Quasi-Newton optimization technique implemented in HCRF toolbox [1].

7 Experiments

We designed our experiments to evaluate our three hypotheses: (1) the effect of modeling semi-structured interviews at the adjacency pair level, (2) the importance of explicitly learning the commonalities between adjacency pairs, and (3) the importance of multimodal features. In this section, we introduce our baseline models and the methodology of our experiments. Furthermore, we compare our model against various baseline approaches.

7.1 Baseline Models

We select two baseline models: (i) a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with a linear kernel, which is widely used as a discriminative model, (ii) a Maximum Entropy Model, which is an energy model similar to the HCRF but without the hidden states assumption. We used MaxEnt models instead of CRF models (Lafferty et al., 2001), as CRFs are designed to predict a sequence of labels while our task contains only one label for the entire interaction.

Support Vector Machine (SVM)
We use the implementation of SVM from the libsvm package (Fan et al., 2008). The parameter that controls the scale of the soft margin was obtained automatically using cross validation. We train two SVM models: one using the averaged features extracted over the entire interview(SVM Holistic), and the second using features from each adjacency pair stacked into a large feature vector (SVM AP).

Maximum Entropy Model (MaxEnt)
MaxEnt is implemented based on Ratnaparkhi (1996). We trained two models: MaxEnt Holistic, MaxEnt AP, following the same technique described for SVM.

7.2 Experiment Settings

All models in this paper were evaluated with the same cross validation and training-testing splits. We use a 4-fold testing and 3-fold validation with retraining. Validation of all model hyperparameters (regularization terms and number of hidden states) was performed automatically. For HCRF, we perform grid search over the regularization constant, 0, 1, 10, 100, 1000, and the number of hidden states, 2, 3, 4, 5. We found the best hyper-parameter setting to be 1 for the regularizer and 4 for the number of hidden states. The reported model parameters are calculated using all available data, with 5-fold cross validation.

We compute precision by taking the number of correctly predicted depressed subjects divided by the total number of subjects that are predicted as depressed. Likewise, recall is computed as the number of correctly predicted depressed subjects divided by the actual number of depressed subjects. The F1 measure is the harmonic mean of the precision and recall in multimodal analysis (Stratou et al., 2013), which is a standard measure to capture the joint performance of precision and recall.

Z-score normalization is performed for each conversation to scale all the features into the same range, making the learned weights comparable. All multimodal features defined in Section 4 are concatenated into one feature vector per observation, in an early fusion fashion. The distribution of depressed and non-depressed subjects is skewed (30 depressed versus 100 non-depressed).

8 Results and Discussion

In this section, we present the results of our three experiments, looking at the effects of adjacency pairs, hidden commonalties and multiple osmolalities of the features. We further analyze the weights learned from our multimodal HCRF model to draw knowledge and implications from our interview corpus.

8.1 Effect of Using Adjacency Pairs

In order to show the benefits of modeling features at adjacency pair level, we compared the holistic approaches (SVM Holistic and MaxEnt Holistic)

---

[1] http://sourceforge.net/projects/hcrf/

Model	F1	Precision	Recall
HCRF	**0.664**	**0.767**	**0.585**
SVM Holistic	0.417	0.500	0.357
SVM AP	0.449	0.533	0.381
MaxEnt Holistic	0.523	0.567	0.486
MaxEnt AP	0.603	0.733	0.512

Table 1: Comparison of our approach with baseline models. 'Holistic' stands for models with features extracted over the whole interaction, 'AP' stands for models with features extracted at adjacency pair level.

with the adjacency pair approaches (SVM AP and MaxEnt AP) by performing pairwise T-tests on a 4 fold testing set. By F1 measure, the adjacency pair approaches are significantly better than holistic approach for both SVM and MaxEnt ($p < .05$ respectively). Detailed numbers are shown in Table 1. This shows that using features extracted at each adjacency pair level is better than extracting features over the whole interaction in the task of depression prediction as we have hypothesized in H1 of Section 3. Extracting features at the entire interview level ignores discriminative information within each adjacency pair as well as the dependence between consecutive pairs.

8.2 Effect of Learning Commonalities among Adjacency Pairs

Multimodal HCRF automatically learns the commonalities among different adjacency pairs by assigning them to the same hidden state. Each hidden states is a similar set of questions designed to serve similar purpose. We see from Table 1 that our approach outperforms all the baselines. Four paired T-tests are performed on the F1 measures, between the HCRF and each baseline model (SVM Holistic, SVM AP, MaxEnt Holistic and MaxEnt AP) on a 4-fold testing set and found statistical significance in all the four pairs with $p < .05$. These results suggest the advantage of learning commonalities among adjacency pairs, as we have hypothesized in H2 of Section 3.

8.3 Effect of Using Features Extracted from Three modalities

Figure 3 shows that the use of features from three modalities statistically outperforms (paired T-test with $p < .05$) all other possible combination of modalities using HCRFs in terms of the F1 measure, as we have hypothesized in H3 of Section 3. These results confirm the advantage of combining features from three modalities in the depression prediction task suggested in our third hypothesis. Yu et al. (2013) reported similar trends in friendship prediction.

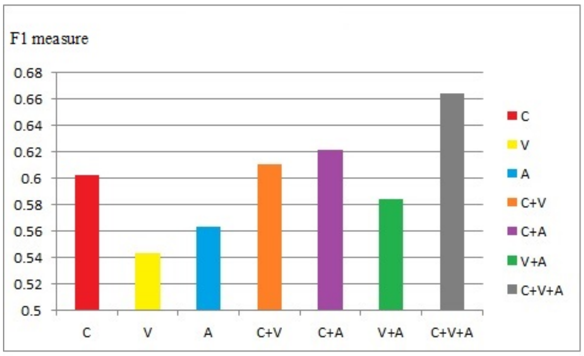


Figure 3: Comparison of our comprehensive multimodal approach against other set of features using HCRF, 'C' stands for conversational features, 'V' stands for visual features and 'A' stands for acoustic features, '+' stands for combination

8.4 Analysis of the Learned Multimodal HCRF

Figure 4 illustrates the learned Multimodal HCRF model with its optimized parameters. The learned model has four hidden states, which means that the adjacency pairs are clustered into four groups. By analyzing θ_y, we observe that depressed individuals are more tightly associated with the verbal and nonverbal behaviors manifested in the first and the last hidden states, while non-depressed individuals are more tightly associated with the second and third hidden states. We obtain the set of the most predictive features for each hidden state by selecting features with associated weights higher than 0.15. For example, in hidden state 1, "speech onset time", "neutral expression", "energy in dB" and "peak slope" stand out as the top ranked features. We show the top ranked features of each hidden state in Figure 4.

By performing inference on the learned model parameters, we can recover a list of the adjacency pairs most strongly associated with each hidden state for each participant. Then we hold a majority vote for each adjacency pair with all 130 participants to determine its most strongly associated hidden state overall. The first hidden state was most strongly associated with the responses to the questions *"How would your best friend describe you?"*, *"Tell me about the last time you felt really happy?"*, and *"I'm sure you can tell by my shoes. I'm not much of a world explorer. Do*

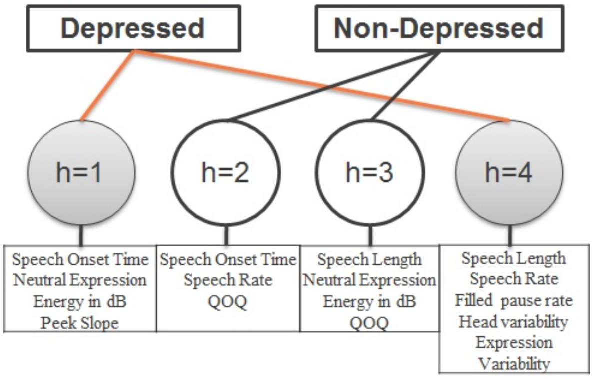


Figure 4: the Multimodal HCRF model for depression prediction. Hidden state 1 and 4 are more correlated with depressed people, while hidden state 2 and 3 have relatively larger influence on non-depressed people. We also listed features with weights higher than 0.15.

you travel a lot?”. It is interesting to see that all of these questions are designed to build up intimacy between clinicians and patients. We found that “speech onset time” is negatively correlated with depression for all three adjacency pairs mentioned above. This is consistent with the findings in Cohn et al. (2009), where increased speaker-switch duration in conversation is found in the depressed group. However, there are other features that are only salient for one adjacency pair but not for the others. For instance, “peak slope” and “energy in dB” are only salient for the first question’s response, but not for the others. The “peak slope” feature has been identified as a good indicator of depression, and as Scherer et al. (2013b) suggests, depressed patients tend to have tighter glottal flow than healthy individuals. Lower “energy in dB”, meaning quieter speech, is correlated with depression. In addition to the above observations, we find that the “neutral expression” feature is not salient. This is despite the feature being the second most heavily weighted feature associated with the first hidden state. We believe that clustering adjacency pairs together through the hidden states provides more predictive power than using the features themselves. A previous study also found that “neutral expression” is a good indicator of depression through a holistic analysis (Stratou et al., 2013).

For the fourth hidden state of our model, the adjacency pairs with questions *“What are things you really like about LA?”*, *“How are you doing?”* , *“Where are you originally from?”*, and *“Sometimes when I’m feeling tense, I turn on the fish tank screen saver. Hey I know it’s not Hawaii but it’s the best I’ve got. What do you do to relax?”* appear to be the most relevant according to majority vote. All of these questions are from the rapport building phase of the interview. We found that for all four questions, depressed participants respond with shorter speech length. This finding is correlated with a previous report that depressed people are less expressive in the rapport-building phase of the conversation (Bylsma et al., 2008). In addition to shorter “speech length”, lower “speech rate” is also a salient indicator of depression in response to the first three adjacency pairs we mentioned above, which correlates with findings of a previous study (Teasdale et al., 1980).

To sum up, our analysis suggests that clinicians should focus on different verbal and nonverbal behaviors in response to different questions. For example, “speech onset time” is very crucial for evaluating responses triggered by intimate questions, while “speech length” is very important for rapport building questions.

9 Conclusion

We introduced the Multimodal HCRF, a computational model which explicitly considers the context and the commonalities among the adjacency pairs in an interview. By combining conversational, visual and acoustic features, our model outperforms the use of any other combination of the modalities. The saliency of the verbal and nonverbal features extracted from the adjacency pairs is related to the content and purpose of the probing questions. For future work, we plan to incorporate linguistic cues, such as sentiment analysis, syntactic structure and lexical features into our computational model.

10 Acknowledgement

We thank the Richard King Mellon Foundation for their generous funding. This work is also partially supported by DARPA under contract (W911NF-04-D-0005) and U.S. Army Research, Development, and Engineering Command. The content of the paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We also thank Alan W Black, Carolyn Rose, Yanchuan Sim, Shoou-I Yu, Elijah Mayfield and Brian Coltin for their insightful suggestions and Ramon Paz, Rob Groome and Ben Farris for their technical support.

References

P. Alku, T. Bäckström, and E. Vilkman. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

T. Baltrusaitis, P. Robinson, and L. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *CVPR, 2012 IEEE Conference*, 2012. doi: 10.1109/CVPR.2012.6247980.

L. Bylsma, B. Morris, and J. Rottenberg. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical psychology review*, 28 (4):676–691, 2008.

J. F. Cohn, T. S. Kruez, I. Matthews, Y. Ying, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, 2009.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

D. DeVault, K. Georgilia, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *to appear in Proceedings of SigDial 2013*, 2013.

T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech 2011*, pages 1973–1976. ISCA, 2011.

H. Ellgring. *Nonverbal communication in depression*. Cambridge University Press, Cambridge, 1989.

L. A. Fairbanks, M. T. McGuire, and C. J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology*, 91(2):109–119, 1982.

R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

H. P. Grice. Logic and conversation. *1975*, pages 41–58, 1975.

J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician-patient interaction. *Applied and Preventive Psychology*, 4(1):21–37, 1995.

N. Henrich, C. d'Alessandro, and B. Doval. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. *Proceedings of EUROSPEECH, Scandanavia*, pages 47–50, 2001.

F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23:67–72, 1975.

J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech 2011*, pages 177–180. ISCA, 2011.

J. Kane, S. Scherer, L.-P. Morency, and C. Gobl. A comparative study of glottal open quotient estimation techniques. In *to appear in Proceedings of Interspeech 2013*. ISCA, 2013.

A. Kirsch and S. Brunnhuber. Facial expression and experience of emotions in psychodynamic interviews with patients with ptsd in comparison to healthy subjects. *Psychopathology*, 40 (5):296–302, 2007.

K. Kroenke and R. L. Spitzer. The phq-9: A new depression and diagnostic severity measure. *Psychiatric Annals*, 32:509–521, 2002.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL `http://dl.acm.org/citation.cfm?id=645530.655813`.

G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *FG 2011, 2011 IEEE International Conference on*, 2011. doi: 10.1109/FG.2011.5771414.

R. Menke. *Examining nonverbal shame markers among post-pregnancy women with maltreatment histories*. PhD thesis, Wayne State University, 2011.

L. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *FG '08. 8th IEEE International*

Conference on, 2008. doi: 10.1109/AFGR.2008.4813429.

J. E. Perez and R. E. Riggio. *Nonverbal social skills and psychopathology*, pages 17–44. Nonverbal behavior in clinical settings. Oxford University Press, 2003.

A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Advances in neural information processing systems*, pages 1097–1104, 2004.

A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1848–1852, 2007.

A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142, 1996.

E. A. Schegloff. *Sequence organization in interaction: Volume 1: A primer in conversation analysis*, volume 1. Cambridge University Press, 2007.

J. T. M. Schelde. Major depression: Behavioral markers of depression and recovery. *The Journal of Nervous and Mental Disease*, 186(3): 133–140, 1998.

S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Proceedings of Interspeech 2013*. ISCA, 2013a.

S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*. IEEE, 2013b.

G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *to appear in Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013.

D. Talkin. A Robust Algorithm for Pitch Tracking. In W. B. Kleijn and K. K. Paliwal, editors, *Speech coding and synthesis*, pages 495–517. Elsevier, 1995.

J. D. Teasdale, S. J. Fogarty, and J. M. G. Williams. Speech rate as a measure of short-term variation in depression. *British Journal of Social and Clinical Psychology*, 19(3):271–278, 1980.

Z. Yu, D. Gerritsen, A. Ogan, A. W. Black, and J. Cassell. Automatic prediction of friendship via multi-model dyadic features. 2013.

Poster Abstracts

Gaze cue effect during language comprehension

Dato Abashidze
Cognitive Interaction Technology Excellence Cluster, Bielefeld University,
Bielefeld, Germany
dabashidze@cit-ec.uni.bielefeld.de

Pia Knoeferle
Cognitive Interaction Technology Excellence Cluster, Bielefeld University,
Bielefeld, Germany
knoeferl@cit-ec.uni-bielefeld.de

Maria Nella Carminati
Cognitive Interaction Technology Excellence Cluster, Bielefeld University, Bielefeld, Germany
mcarmina@techfak.uni-bielefeld.de

1 Introduction

Real-world eye-tracking results from 3 experiments suggest that people prefer to look at recently depicted over possible future events during spoken sentence comprehension. Participants (*N*=32) saw a videotaped actor performing an action (e.g. sugaring strawberries). Once the action was completed, they heard a German sentence (NP1-VERB-ADVERB-NP2) that referred either to that action (e.g., *Der Versuchsleiter zuckerte kürzlich die Erdbeeren* 'The experimenter recently sugared the strawberries', or an equally plausible action that the actor would perform in the near future (e.g., *Der Versuchsleiter zuckert demnächst die Pfannkuchen* 'The experimenter will soon sugar the pancakes'). People's eye movements to the objects were recorded while they heard the sentence (Fig. 1).

In Expt 1 by Knoeferle et al., (2011, Expt 1) participants only saw the past action being performed (see also Abashidze et al., 2011, Exp. 1; Knoeferle & Crocker, 2007, Exp. 3). The results showed that, although at the ADVERB the sentence becomes fully disambiguated towards the past or the future event, looks to the past (strawberries) and future (pancakes) objects only started to diverge late during the NP2 (i.e., the tense effect). Until then, listeners preferred to look at the recent object (strawberries). Crucially, throughout the sentence there was an overall preference to look at the past than the future object, irrespective of sentence tense (henceforth 'recent-event preference'). This recent-event preference was investigated further in two subsequent experiments. In Expt 2 by Knoeferle et al. (2011), participants saw both the past and future action performed equally often (50% frequency), while in Expt 3 (Abashidze et al., 2013) the frequency of the future action was increased to 75% of the trials (vs. 25% for the past action). As a result of these frequency manipulations, looks to the past and future object started to diverge earlier - in the later part of the ADVERB region in Expt 2, and at the end of the verb region in Expt 3. However, the overall bias of looking more at the past vs. future object remained present throughout most of the sentence. Clearly, the recent-event inspection preference is robust and not easily overridden by frequency manipulations favoring a future event.

In our latest study the recent-event preference was pitted against a situational cue that seems to be very effective in directing visual attention, i.e. gaze. Gaze is important in communication and existing research has examined how a listener responds to a speaker's gaze during language comprehension. A study by Hanna and Brennan (2007) examined gaze cues in speaker/listener pairs during a simple target-matching task. They found that listeners used the gaze cues of speakers to identify correct targets before the point of linguistic disambiguation. A study by Knoeferle and Kreysa (2012) examined effects of a speaker's gaze on a listener's visual attention and language comprehension when the speaker did not directly face the listener. The results showed that even when the speaker was positioned at an angle to the listener, the listener followed the speaker's gaze to the target objet before it was mentioned (see also Macdonald, & Tatler 2013).

With regard to the recent event preference and our experiments, we wanted to see whether and to which extent an actor's gaze towards the (past or future) object influences listeners' visual attention; in particular, we wanted to see whether a gaze towards the future object could overcome the preference for the recently acted upon object. The current study (*N*=32) used the same experimental materials as the previous studies (e.g., Knoeferle et al., 2011). In addition we created short 'gaze' video clips for every item, showing the experimenter gazing at the target object (e.g. past (strawberries) or future object (pancakes)).

As in the previous studies, the videotaped experimenter performed one action before the sentence – the recent action (e.g., sugaring the strawberries) and then after 700 ms the sentence was presented. In half of the trials the experimenter gazed at the target object from VERB onset and kept his gaze on the target until the end of the sentence. In the other half of the trials, participants saw a static picture of the experimenter looking straight ahead. The second (i.e. future) action was shown 700 ms after the end of the sentence. Thus, the experiment manipulated 2 factors: sentence tense (past vs future) and gaze to target object (gaze vs no gaze). Past and future events were shown equally often.

Fig.1 Example of experiment

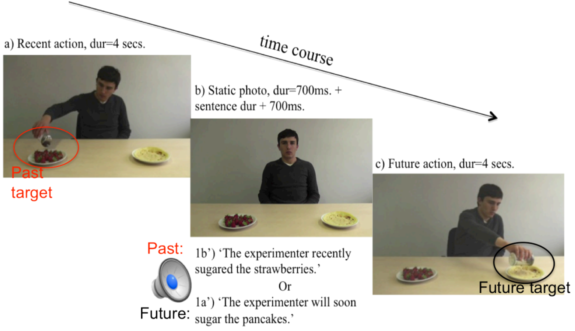

time course
a) Recent action, dur=4 secs.
Past target
b) Static photo, dur=700ms. + sentence dur + 700ms.
c) Future action, dur=4 secs.
Future target
Past:
1b') 'The experimenter recently sugared the strawberries.'
Or
Future:
1a') 'The experimenter will soon sugar the pancakes.'

Fig. 2 Mean log gaze probability ratios (*ln*(*P*(recent target)/*P*(future target)) from verb onset

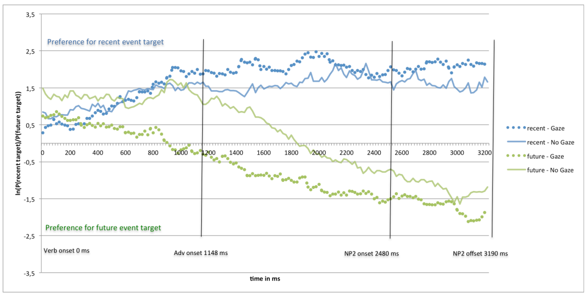


Fig. 2 shows the time course of participants' eye fixations from verb onset. The dependent measure is the mean log gaze probability ratio (*ln*(*P*(recent target)/*P*(future target)).This ratio expresses the visual bias strength for the past target vs. the future one. A positive value means more looks to the past target, a negative one to the future. In the no-gaze conditions, we replicated the results of Expt 2 (Knoeferle et al., 2011), with the preference for looking at the past object reversing only during the adverb (where ratio values become negative). When gaze (vs. no gaze) was available, there were more and earlier looks to the target object. Importantly, gaze affected looks to the future object to a greater extent than the past object. With gaze (cf. green dotted line, Fig 2), looks to the future object increased faster than without gaze (solid green line), with the ratio becoming negative (showing a preference for the future target) 800ms earlier than with no gaze (1100 vs. 1900ms). In sum, by triggering more and earlier looks to the future target, gaze mitigated the recent event preference, however, it did not completely override it, as in the first 800 ms there was still an overall preference for the past target.

References

Abashidze, D., Knoeferle, P., & Carminati, M.N. (2013). Do comprehenders prefer to rely on recent events even when future events are more likely to be mentioned? Poster presented at *AMLaP* 2013 in Marseille, France.

Hanna, J. E. & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57, 596-615.

Knoeferle, P., & Crocker, M.W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye movement. *Journal of Memory and Language* 57(4), 519–543.

Knoeferle, P., Carminati, M., Abashidze, D., & Essig, K. (2011). Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension. *Front. Psychology* 2:376. doi: 10.3389/fpsyg.2011.00376.

Knoeferle, P., & Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? *Frontiers in Psychology* 3, 538.

Macdonald, R. G., & Tatler, B. W. (2013). Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of Vision*, 13 (4):6, 1-12.

Conversations and Incomplete Knowledge

Nicholas Asher[1] **and Soumya Paul**[2]
[1]CNRS, IRIT and [2]National University of Singapore

1 Introduction

Conversations often involve an element of planning and calculation of what to say to best achieve one's interests. We investigate scenarios of incomplete knowledge in strategic conversations, where the fundamental interests of the dialogue agents are opposed. For instance, a debate between two political candidates. Each candidate has a certain number of points she wants to convey to the audience, and each wants to promote her own position to the expense of the other's. To achieve these goals each participant needs to plan for anticipated responses from the other. Debates are thus games; an agent may win, lose or draw. Similar strategic reasoning about what one says is a staple of board room or faculty meetings, bargaining sessions, etc. We show the importance of a certain form of unawareness in strategic conversation.

We explore a linguistic consequence of the model of strategic conversation of (Asher and Paul, 2012; Asher and Paul, 2013) concerning a form of incomplete information, where one strategic player is unaware of moves that another player may perform. We show some interesting linguistic consequences of the model concerning this form of incomplete information and draw an abstract characterization of the structure of strategic conversations from the framework. This work complements more computational and empirical work like that of (Traum, 2008).

Background. For their model of strategic conversations, (Asher and Paul, 2012) use Banach Mazur or BM games, a kind of infinitary game (Kechris, 1995) used in mathematics and theoretical computer science. A BM game is a win-lose game $\langle X^\omega, \mathit{Win}\rangle$ involving two players; the 2 players each play a finite sequence of moves from a fixed set or vocabulary X, alternating indefinitely and building strings in X^ω; $\mathit{Win} \subseteq X^\omega$ is the winning condition for player 0 (for player 1 the winning condition is $X^\omega - \mathit{Win}$). The Cantor topology over X^ω of infinite strings allows us to characterize winning conditions in terms of basic open sets, unions of basic open sets (Σ^0_1), intersections of complements of basic open sets (Π^0_1), and so on. The *Borel hierarchy* consists of the Σ^0_1 sets, the Π^0_1 sets, and more generally includes $\Sigma^0_{\alpha+1}$ as the countable union of all Π^0_α sets and $\Pi^0_{\alpha+1}$ as the complement of $\Sigma^0_{\alpha+1}$ sets. The hierarchy is strict and does not collapse (Kechris, 1995).

(Asher and Paul, 2012) characterize types of dialogues and their conversational goals using the BM framework. Message exchange games are BM games $\langle X^\omega, \mathit{Win}\rangle$ where X is a set of possible discourse moves, as described by, e.g., SDRT (Asher and Lascarides, 2003). BM games characterize in a precise way how some conversational strategies, and some winning conditions in strategic conversations, are much more complex than others. (Asher and Paul, 2012) also show how two conversationalists 0 and 1 may occupy a role in two different BM games such that 0 and 1 may both have winning strategies (1 in each game) and how this applies to cases of misdirection (Asher and Lascarides, 2013). Finally, BM games also can model why speakers do not "defect" when given the opportunity and it is in their interest. Consider a prosecutor who asks a defendant a question that may incriminate her and that she prefers not to answer. In a one shot linguistic exchange, it is not rational to answer such a question. However, if linguistic games are open ended allowing for further exchanges, then a defection strategy may carry heavy, known penalties.

Our contribution. BM games are determined (Martin, 1975); so if 0 and 1 are playing a game G in which each has complete common knowledge of the moves and strategies of the other, it is *not* rational for both 0 and 1 to play with a strict preference for winning. If they do play with such a preference, they must *not* have common and complete knowledge of the game they are playing. We investigate two scenarios of incomplete and non-common knowledge: one is where the players are playing with different sets of moves and so the moves of one are not completely known to the other; the other is where players start out with the same repertoire of moves, but one forgets (or learns) certain moves and the other does not. In both scenarios the players are playing different games G and G' with sets of moves X and Y respectively such that $X \subsetneq Y$. In this case, one player will be unaware of some of the moves available to the other.

A question then is: if player 0 strategizes for $Win\ \phi$, what happens to ϕ in the game where player 1 has a set of moves available to him that is a strict superset of those 0 is aware of? (Asher and Paul, 2013) prove an abstract result showing that Win_X encoded in G' may have a higher Borel complexity. For our part we are just interested in the restriction of the theorem that states that a winning condition that has complexity Σ^0_1 in G will jump to Σ^0_2 in G'.

We illustrate the theorem's import with an excerpt from the Dan Quayle-Lloyd Bentsen Vice-Presidential debate of 1988. Quayle, as a very junior and politically inexperienced Vice-Presidential candidate, was repeatedly questioned about his experience and his qualifications to be President. Quayle's strategy to rebut doubts about his qualifications was to compare his experience to the young John Kennedy's. However, Bentsen made a discourse move that Quayle didn't anticipate.

(1) Quayle: ... the question you're asking is, "What kind of qualifications does Dan Quayle have to be president," [...] I have as much experience in the Congress as Jack Kennedy did when he sought the presidency.

(2) Bensten: Senator, I served with Jack Kennedy. I knew Jack Kennedy. Jack Kennedy was a friend of mine. Senator, you're no Jack Kennedy.

Quayle's strategy at that point fell apart, and he lost the debate handily. He was unprepared for Bentsen's move, which we model by having Quayle play a game with set of moves X and Bentsen a game with set of moves Y such that $X \subsetneq Y$.

The theorem implies that a winning strategy for Quayle's winning condition—implicating that he was comparable to a very distinguished President (a Σ_1 winning condition)—would have needed to take into account an intersection of open sets in Y defining the X winning condition in Y thus anticipating possible deviations from the conversational plays in X. Had he done so, he might have countered Bentsen's move and have kept the moves within X. A linguistic theory of discourse structure like SDRT tells us how:

Proposition 1

If a move α presupposes ϕ and ϕ is not locally accommodatable in α and a move β is such that $\beta \models \neg\phi$, then there is no link between α and β. I.e. α cannot be a response to β.

In this case, Bentsen's move presupposes that Quayle had implicated or said that he was comparable to John Kennedy, a presupposition that is not locally acommodatable (to Bentsen's move). Had Quayle explicitly added a rider to his response, like *though I would not presume to be the great statesman that Kennedy was, I have as much experience as he did when he sought the presidency*, Bentsen's move would have been incoherent and would have put him in a position to lose the debate.

BM games offer a simple and elegant way of describing a heretofore little studied form of unawareness, an unawareness of moves in the game instead of an unawareness of events (Haifetz et al., 2006). It is the former that is appropriate for the analysis of strategic conversation. Our observations provide a general characterization of the structure of strategic conversations, assuming that our dialogue agents are rational and are perfect reasoners, thus able to determine whether a winning strategy exists in the game they are playing.

Proposition 2 *Two rational players of a BM message exchange game assign a strict preference to their winning conditions only if they (i) are playing two games with compatible Win conditions, or (ii) assume they are playing a game where their opponent is unaware of some of their moves.*

Case (i) is the misdirection scenario; case (ii) includes both cases of forgetting and of assuming your opponent doesn't know all of your rhetorical repertoire. The result resembles exceptions due to unawareness of no speculative trade theorems in economics (Milgrom and Stokey, 1982).

References

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

N. Asher and A. Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6(2):http://dx.doi.org/10.3765/sp.6.2.

N. Asher and S. Paul. 2012. Conversations as Banach-Mazur games. *Dialogue and Discourse (submitted)*.

N. Asher and S. Paul. 2013. Infinite games with uncertain moves. In *Proceedings of the First International Workshop on Strategic Reasoning*, pages 25–32, Rome.

A. Haifetz, M. Meier, and B. Schipper. 2006. Interactive unawareness. *Journal of Economic Theory*, 130:78–94.

A. Kechris. 1995. *Classical Descriptive Set Theory*. New York: Springer-Verlag.

D. A. Martin. 1975. Borel determinacy. *Annals of Mathematics*, 102(2):363–371.

Paul Milgrom and Nancy Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory*, 26(1):17–27.

D. Traum. 2008. Computational models of non-cooperative dialogue. In *Proceedings of the International Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL)*, London.

Turn-timing in naturalistic mother-child interactions: A longitudinal perspective

Marisa Casillas[a,c], Elma Hilbrink[a,d], Susan C. Bobb[b], Eve V. Clark[c],
Merideth Gattis[d], & Stephen C. Levinson[a]
a. MPI Nijmegen for Psycholinguistics,
b. Northwestern University, c. Stanford University, d. Cardiff University
Corresponding email: `Marisa.Casillas@mpi.nl`

Abstract

Combining data from two longitudinal studies of young children, we track the development of turn-timing in spontaneous infant-caregiver interactions. We focus on three aspects of timing: overlap, gap, and delay marking. We find evidence for early development of turn-timing skills, in-line with the Interaction Engine Hypothesis.

Part and parcel of learning a language is learning how to *use* it. Conversation is our first and primary mode of language use, and determines the form of children's linguistic input. But participating in conversation is not trivial; it requires interactants to weave together linguistic, non-verbal, and interactional information in real time, both while speaking and listening. Places of turn transition—when one speaker stops and the next one can start—are especially difficult since the interactants must coordinate on who will speak next, and when. Nonetheless, adults manage to take turns with apparent ease; their turn-transitions occur with minimal vocal overlap and gap between spoken turns. When speakers can't respond with immediate timing, they often delay their turn with markers such as *uh* and *um*. These patterns of turn-timing have been demonstrated in cultures around the globe, and thus appear to undergird human conversation (Clark & Fox Tree, 2002; Stivers et al., 2009). Cross-cultural universals in interactive structure are predicted by the Interaction Engine hypothesis, which suggests that human interactive abilities developed earlier and independently from linguistic abilities (IEH; Levinson, 2006). Applied to infant development, this same hypothesis predicts that infants begin to master interactive skills early and independently from their linguistic skills. We tested this idea by analyzing turn-timing in spontaneous interactions between English-speaking mothers and their children from 0;3 to 3;4. In-line with the IEH, we found that three aspects of turn-timing—vocal overlap avoidance, silent gap minimization, and marking response delays—emerge early in development and interact with children's linguistic planning once they begin to speak.

We analyzed turn-timing in two longitudinal free play corpora: (C1) 10-min in-lab recordings for 12 infants at 0;3, 0;4, 0;5, 1;0, and 1;6 (Ellis-Davies et al., 2012), and (C2) 1-hour at-home recordings for 5 children at 1;8, 2;0, 2;4, 2;8, 3;0 and 3;4 (Demuth, Culbertson, & Alter, 2006). In the first corpus we measured the timing of all transitions between vocalizations by the mother and baby.[1] In the second corpus, we measured the timing between 30 questions and answers for each child at each time point, and further coded each response for its complexity and markers of delay. We also measured the silent gap following turn-initial delay markers and preceding the rest of the turn (e.g., the '..' in "um .. that one"). As is typical of infant-parent interaction (Henning, Striano, & Lieven, 2005), most of the transitions from mother to baby in the first corpus were formatted as questions or ended in tag questions. Thus, the data from both corpora primarily represent turn-timing behavior in question-answer pairs.

Children and mothers took turns vocalizing throughout our sample (0;3–3;4). But, before 0;5, children frequently came in too early; they overlapped their vocalizations with the end of their mothers' nearly 40% of the time. At 0;5, children's overlaps began to decrease, matching the mothers by 1;6, and falling below them, to approximately 4%, by the first sample of the second corpus at 1;8 (Figure 1). This may suggest that children begin to avoid overlap at 0;5, respecting the norm of "one speaker at a time" (Sacks, Schegloff, & Jefferson, 1974).

---

[1] Except transitions from mother to baby when the baby's turn constituted a burp, sneeze, cough, etc.

Because of children's frequent overlap early on, their average turn-timing appears almost adult-like. Quick turn-timing in the first five months has also been reported in prior work (e.g., Ginsburg & Kilbourne, 1988), however these results are likely due to children's high frequency of overlapped starts during this early period. If we instead look at children's *gaps* (non-overlapped starts) with time, we see a clearer picture. Children start out on par with their mothers, but show significantly longer gaps at the 12-month sample. This increase in gap duration slowly tapers off over the rest of the sample until children converge with their mothers' timing again at 2;8 (Figure 2). The non-linear trajectory of gap timing (i.e., rise-then-fall) peaks near the onset of children's first words. If children's slower timing were really due to linguistic planning, we should find that more complex responses have longer gaps than less complex ones. We confirmed this with a linear mixed effects model of turn-timing in the second corpus, finding that more complex answers yielded longer gap durations (yes-no vs. wh-, single nominal vs. inflected phrase, $p<.001$) for children's, but not mothers' answers. This suggests that children may begin to minimize their gaps in the first year, but that the onset of speech may create significant planning costs and disrupts their ability to give an immediate response. Because of this, it may become crucial for children to mark their delays in speaking after 1;0.

Turn-initial delay markers (e.g., *uh*, *um*, prolongation, and repetition) emerged by 2;0 for all five children in the second corpus. Turns beginning with delay markers had significantly more linguistic material than those without, suggesting that children used delay markers when planning more complex responses ($p<.01$). Delay-marked responses were more complex, and so should have shown slower timing overall. However, the delay markers acted to buffer children's extra planning costs effectively, so that turns beginning with delay markers were not significantly longer than those without. Finally, by 3;6, children began to mark delay differentially, just like adults do, using *um* for longer delays and more difficult planning compared to *uh* (Clark & FoxTree, 2002, see also Hudson Kam and Edwards, 2008; Table 1). Children's delay marking suggests that they are cued in to the temporal sensitivity of transferring the floor from one speaker to the next.

In sum, we find that three aspects of turn-timing—overlap, gap, and delay marking—emerge early in children's development. Overlap is acquired first, with children making strides toward adult-like overlap patterns at 0;5. Gaps, too, appear short in the first year, but the onset of speech may cause children to slow down before they improve their overall timing. After the onset of speech, children begin to mark delay, holding the floor when planning complex responses and indicating their attention to the sensitivity of turn transitions. In-line with the IEH, the longitudinal outlook from these data support the idea that turn-timing skills develop early and independently from language, but also are consistent with the fact that, once children begin to speak, the linguistic and interactional systems must converge for children to continue developing adult-like conversational behavior.

References

H. H. Clark and J. E. Fox Tree. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111, 2002.

K. Demuth, J. Culbertson, and J. Alter. Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, 49(2):137–173, 2006.

K. Ellis-Davies, E. Sakkalou, N. C. Fowler, E. E. Hilbrink, and M. Gattis. CUE: The continuous unified electronic diary method. *Behavior Research Methods*, 44(4), 1063–1078, 2012.

G. P. Ginsburg and B. K. Kilbourne. Emergence of vocal alternation in mother-infant interchanges. *Journal of Child Language*, 15, 221–235, 1988.

A. Henning, T. Striano, and E. V. Lieven. Maternal speech to infants at 1 and 3 months of age. *Infant Behavior and Development*, 28, 519–536, 2005.

C. L. Hudson Kam and N. A. Edwards. The use of uh and um by 3-and 4-year-old native English-speaking children: Not quite right but not completely wrong. *First Language*, 28(3):313–327, 2008.

S. C. Levinson. On the human 'interaction engine'. In N. J. Enfield and S. C. Levinson, editors, *Roots of human sociality: Culture, cognition and interaction*, pages 279–298. Berg, Oxford, England, 2006.

H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematic for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.

T. Stivers, N. J. Enfield, P. Brown, ... and S.C. Levinson. Universals and cultural variation in turn-taking in conversation. *PNAS*, 106:10587–10592, 2009.

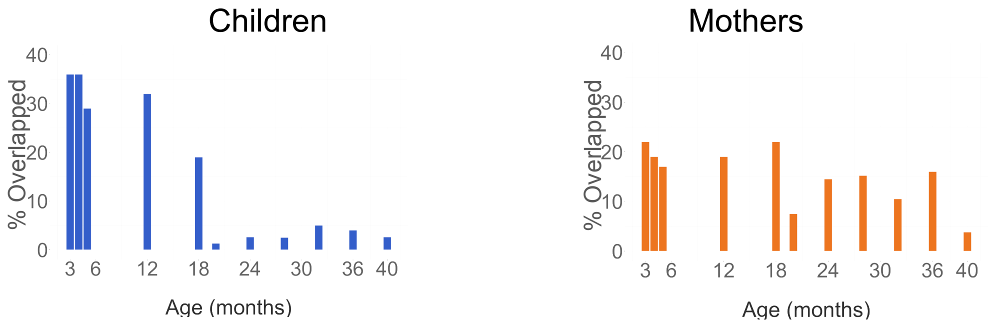


Figure 1. Percent overlapped vocalizations by children (left) and mothers (right) by children's age

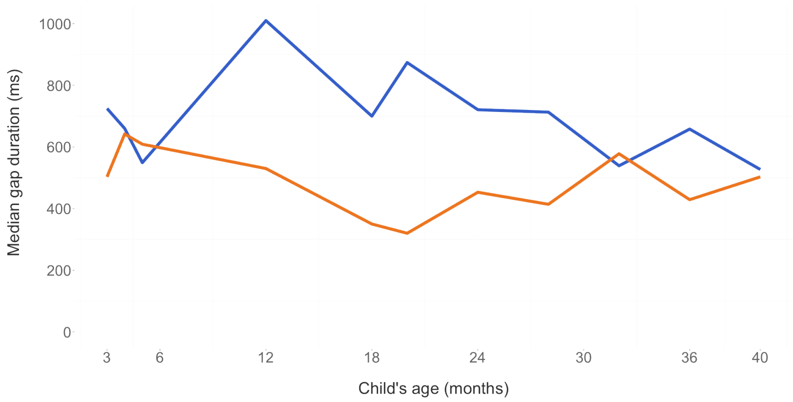
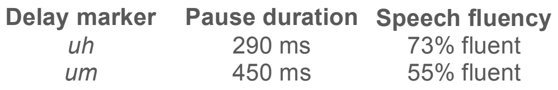


Figure 2. Median gap duration for children (top) and mothers (bottom) by children's age.



Delay marker	Pause duration	Speech fluency
uh	290 ms	73% fluent
um	450 ms	55% fluent

Table 1. Pause duration and speech fluency following children's use of *uh* and *um*.

Putting secondary propositional meanings back on the table

Elena Castroviejo
Spanish National
Research Council
elena.castroviejo@cchs.csic.es

Laia Mayol
Universitat Pompeu Fabra
laia.mayol@upf.edu

E. Allyn Smith
Université du Québec à Montréal
smith.eallyn@uqam.ca

1 Introduction[1]

This contribution examines the felicity of various two-turn dialogues of non-acceptance (DNA), where an initial sentence is followed by a direct *No* or *No that's not true* response. We present experimental results for three languages as well as a modified discourse model and theory accounting for these results.

2 Background

Refutation is often used as a diagnostic, including the direct form as a diagnostic for at-issueness (Simons et al., 2011) and indirect forms as a diagnostic for projective meanings or presuppositions (cf. the *Hey, wait a minute!* test in von Fintel 2004). The motivation behind each of these is that *No* is felicitous with primary asserted content and infelicitous with secondary content (either presuppositions or/and various kinds of implicature). Thus, we would expect to see a pattern like the following, where direct refutation of the not-at-issue presupposition of *again* is not acceptable.

A. John is at the zoo again.
B. No, he's home sick.
B. #No, he's never been to the zoo until now.

A number of theoretical discourse models make the same prediction and/or attempt to explain why we see this pattern, such as Anderbois et al. (2011), Farkas and Bruce (2010), and Schlenker (2012). Farkas and Bruce (and others, following them), posit that primary or at-issue meanings propose a proposition, which leaves room for negotiation, while secondary appositive meanings impose a meaning directly onto the common ground, not allowing for a traditional direct refutation. (Here, appositives, along with expressives, are considered Conventional Implicatures: CIs). And since presuppositions are supposed to already be in the common ground prior to utterance, it is only the primary meaning that ends up on what they call *the Table* to be accepted or refuted. The Table is similar to the stack of topics or questions under discussion, but differs in ways that are not pertinent here. Another model, that of van der Sandt and Maier (2003), makes the opposite prediction that every meaning type should be able to be denied at least in some contexts, and their theory elegantly derives the different possible intended negations of the primary assertion, presupposition, etc. Neither of these approaches can account for data showing that some secondary meanings are more easily denied than others, which is what we find in each language we test (below).

3 Experiments

200+ participants who were native speakers of English, Spanish or Catalan listened to 88 two-turn dialogues (majority fillers) across 4 conditions depending on the type of direct refutation:

1. No, that's not true. $\neg$ p. [NTNT-neg.]
2. No, $\neg$ p. [NO-negation]
3. No, that's not true; q. [NTNT-alt.]
4. No, q. [NO-alternative]

In the first turn of each DNA, a statement was made that crucially contained one of 6 meaning types or subtypes: primary assertion, presupposition: lexical trigger (iterative), presupposition: cleft, CI: appositive, CI: referential expressive, or CI: emotive expressive. The experiment was a Likert judgment task, where participants needed to rate how strange would it be – if strange at all – to overhear someone utter a specific response to the initial sentence. In the figures that follow, the y-axis shows the felicity rating (higher = more felicitous) of the refutation of the meaning type on

---

[1] Authors appear alphabetically, having made equal contributions.

the x-axis; the 6 (sub)types are in the order listed above, as are the four bars for each subtype representing the four types of direct negation.

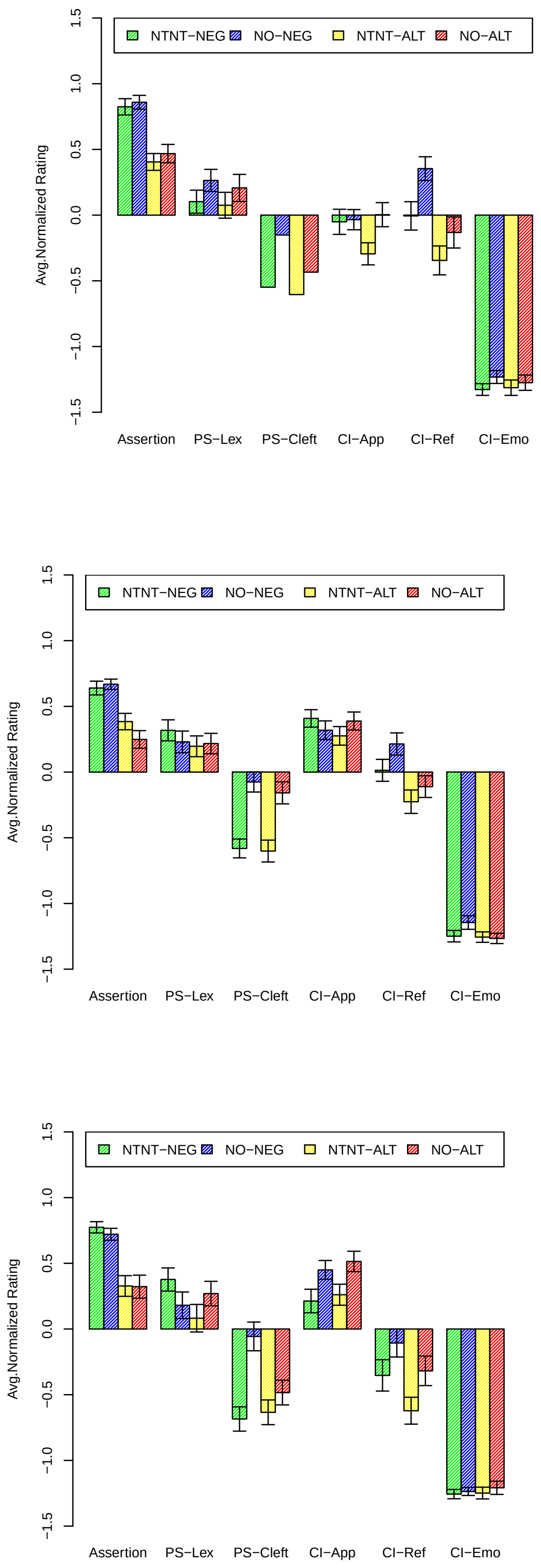


Figure 1: English, Catalan and Spanish results

4 Proposal

What we see across these languages is that, despite a few differences, there is a consistent ranking as follows:

> assertions $>$ referential expressives, appositives $>$ iterative lexical triggers $>$ clefts $>$ emotive expressives

Generally, then, assertions, referential expressives and appositives can be denied more or less felicitously, while clefts and emotive expressives cannot, with lexical triggers somewhere in between. Thus, we need to account for significant differences between different CI types (appositive and ref. v. emotive expressives) and different presupposition types (iteratives v. clefts) as well as explaining why any of them can be directly denied.

What property determines when a secondary meaning will be put on the Table for negotiation (and thus, become a target for direct refutation)? We propose that only those presuppositions or CIs that are propositional and whose propositional content predicates something of an individual are capable of being put back on the Table. By propositional, we mean something of type $< s,t >$ and not the 'expressive propositions' of, e.g. Gutzmann (2011). This accounts for all of the data above and makes additional predictions as well. Emotive expressives are infelicitous because they are non-propositional. In a cleft sentence like *It was John who broke the vase*, the presupposition is either 'someone broke the vase' or 'there is a broken vase', but either way, we don't have something predicated of a specific individual. This predicts that existence presuppositions in general will be infelicitous with direct refutation, which matches our intuition for the case of the definite article (not tested here). Assertions are not subject to this condition since they are already on the Table to begin with, but referential expressives like 'the idiot' and appositives like 'John, an American, ...' both retrieve a referent and predicate something of it, making them deniable. They are also propositional (e.g. 'John is an idiot'), fulfilling both conditions. This theory also hints at a reason iterative triggers may be intermediate in that they are somewhat existential, but it is the existence of an event, which is then predicated of an individual, borrowing something from each side. Thus, our results highlight another variable important for modelling discourses involving varied meaning types.

References

Scott AnderBois, Adrian Brasoveanu and Robert Henderson. 2011. Crossing the appositive/at-issue meaning boundary. *Proceedings of SALT*, volume 20: 328–346.

Donka F. Farkas and Kim B. Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics*, volume 27: 81–118. Oxford University Press.

Kai von Fintel. 2004. Would you believe it? The king of France is back! Presuppositions and truth-value intuitions. In M. Reimer and A. Bezuidenhout, eds. *Descriptions and beyond.* Oxford University Press.

Daniel Gutzmann. 2011. Expressive modifiers and mixed expressives. *Empirical issues in syntax and semantics*, volume 8: 123–141.

Rob van der Sandt and Emar Maier. 2003. *Denials in discourse*. Michigan linguistics and philosophy workshop (ms., University of Nijmegen)

Philippe Schlenker. 2012. *Supplements without bidimensionalism.* (ms., Institut Jean-Nicod and New York University)

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2011. What projects and why. *Proceedings of SALT*, volume 20: 309–327.

Investigating the role of priming and alignment of perspective in dialogue

Simon Dobnik and **Christos Koniaris**
University of Gothenburg, Sweden
`{simon.dobnik,christos.koniaris}@gu.se`

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

Abstract

We examine the alignment of the primed frame of reference (FoR) for spatial descriptions over several utterances of a situated dialogue. We confirm the tendency of FoR alignment and that the *intrinsic* FoR is the most popular one independent of the priming.

1 Introduction

Typically, speakers use projective spatial descriptions such as “to the left of” or “behind” without a specification of the frame of reference (FoR) or perspective according to which the hearers should interpret the scene. For example, they can be interpreted relative to any of the discourse agents or the reference object itself if it can ground orientation of the FoR. The latter is a contextual variable which must be resolved from the visual or discourse context in order for the description to be grounded properly. The resolution of the FoR from the visual context succeeds if a given description can be unambiguously satisfied in it, i.e., a given pair of objects and the relation can be interpreted only according to that perspective for the utterance to be true. FoR can be resolved explicitly by linguistic discourse if the speaker describes it, or if, as we argue in this paper, the perspective is primed (combining observations from both perceptual and linguistic discourse) and aligned over several utterances of a situated conversation. (Carlson-Radvansky and Logan, 1997) is an early example of experimental research that examined the influence of FoR ambiguity on spatial term semantics. A finding from this study was that for vertically aligned prepositions, e.g., *above*, there was a preference for the *hearer/viewer-centric* FoR. Later work, (Kelleher and Costello, 2005) examined the impact of FoR ambiguity with respect to horizontally aligned prepositions, e.g., *in front of*. Interesting, this research reported a preference for the *intrinsic* FoR. Neither of these studies explicitly considered the effect of priming on FoR selection. More recently, (Li et al., 2011) studied the impact of FoR preference on object selection from an array of objects. The results from this study indicate that the intrinsic FoR of the object array was preferred. Again the study did not examine any priming effects. Finally, (Duran et al., 2011) examined the effect of social factors – such as the presence of a social partner and their ability to use a FoR – on reference frame selection. In contrast with the previous studies where the *intrinsic* FoR was preferred, this study found that participants invested in either an other-centric (*speaker-relative* in the terminology used in this paper) or egocentric (*hearer-relative*) mode of responding. In this study we are interested in mechanisms of such priming and alignment of FoR over several utterances in a way that they could be implemented as a model of a dialogue manager of a situated conversational agent (Trafton et al., 2005).

2 Experiment

(Dobnik, 2012) identifies the strategies of reference alignment and coordination by examining a small corpus of situated conversations between two human agents. Here, we build on this work by constraining the scenarios in such a way so that we can study under what conditions the identified strategies are applied and how are they followed. We replace one of the conversational partners with a pre-scripted virtual agent and restrict her utterances to particular scene configurations. In each turn, the agent generates a spatial description of a scene from which the hearer may or may not resolve the reference frame. The human must click on the object referred to by the description and so confirms their interpretation. During the priming step a description and a scene are chosen so that

"I chose the blue box to the left of the chair."

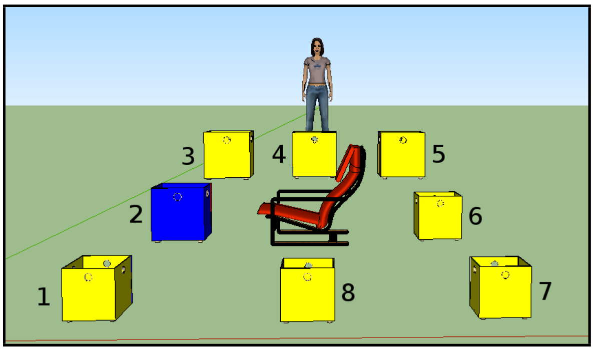


Figure 1: The description and the scene uniquely ground the reference frame in the priming turn of the conversation to H. If boxes 6 and 8 were also blue, then the description would also be interpretable under S and I reference frames. Object numbers were hidden from participants.

only one object matches as the description's reference and hence the description can only be interpreted according to one perspective (speaker-relative (S), hearer-relative (H) and intrinsic (I)) as shown in Figure 1. In the second turn, the system generates another description but in this case it matches the scene ambiguously in respect to all three reference frames, i.e., there would be three objects matching the description, one for each FoR interpretation. The human now has a choice to follow the primed FoR or choose a different one. Following a successful interpretation, the system generates yet another similar description. Finally, in the fourth turn the floor is handed to the human and they are invited to describe the location of an object indicated by the system. The purpose of this turn is to see whether the priming would also be preserved when the speaker–hearer roles change.

Since priming is given for all three FoRs and there are dialogue segments of 4 turns, this gives us "conversations" that contain totally 12 turns per participant (75 participants, 51 complete trials used). In this paper we concentrate only on the results from the second and third turn of each primed dialogue segment (3 × 2 × 51 = 306 utterances). The experiment was implemented as a web page and the results are from both supervised lab sessions and anonymous online contributions.

	Hearer (H)	Intrinsic (I)	Speaker (S)
Primed with	102	102	102
Used by hearer	74	157	75
Priming succeeded	52	78	43
Priming failed	50	24	59
H priming followed by	–	32	18
I priming followed by	10	–	14
S priming followed by	12	47	–
Used instead of prime	22	79	32

Table 1: Summary of the number of utterances/trials according to the FoR assignment

3 Results and discussion

Our findings are presented in Table 1. The first row shows the number of trials over all participants for each of the three primed turns. The second row shows the hearers' own preference for FoR. The third gives us the number of trials where the priming succeeded into alignment and the fourth shows utterances where although a FoR was primed in the previous utterance it was not applied by the hearer in this utterance. The rows 58 list the breakdown of FoR usage for the cases where the user adopted a different FoR to the primed one. Finally, the last row shows the number of utterances a particular FoR was used instead of the primed one.

The results in Row 3 show that priming has an effect on the choice of the FoR in the subsequent utterances (baseline count per primed FoR is 102/3 = 34). The results also show that there is a clear preference for using intrinsic FoR as shown previously in the literature (see Section 1): Row 2 and the breakdown of non-alignment in rows 5–8. Our impression from short discussions with some of the participants is that this choice may be seen as a convenient way of setting a "neutral", objective reference that both the hearer and the speaker can easily refer to during their communication. Furthermore, the chair which is setting the intrinsic FoR is also providing additional *perceptual priming* which may be another contributing factor: the chair is large, red and in the centre of the scene.

4 Future work

In the immediate future work, we will analyse the distribution of alignment between the Turn 2 and 3, and the behaviour of the hearer when they take on the role of a speaker (data from Turn 4) and also extend our experiments to include conditions under which reference objects receive different visual priming.

References

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.

Simon Dobnik. 2012. Coordinating spatial perspective in discourse. In Pierre Nugues, editor, *Proceedings of the 4th Swedish Language Technology Conference (SLTC 2012)*, pages 21–22, Lund, October 24–26.

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

John Kelleher and F Costello. 2005. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-Sigsem Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms and Applications*.

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.

J Gregory Trafton, Nicholas L Cassimatis, Magdalena D Bugajska, Derek P Brock, Farilee E Mintz, and Alan C Schultz. 2005. Enabling effective human-robot interaction using perspective-taking in robots. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 35(4):460–470.

Toward a Tutorial Dialogue System for Urban Navigation

Christopher Fox and Kristy Elizabeth Boyer
Department of Computer Science
North Carolina State University
Raleigh, North Carolina, USA
{cdfox,keboyer}@ncsu.edu

GPS navigation systems are tremendously beneficial for drivers, affording their users the ability to navigate to any destination regardless of their prior knowledge of the roads. However, as previous studies have shown, these systems can inhibit the formation of a *cognitive map* of a driver's local area (Jackson, 1996; Burnett and Lee, 2005). This lack of development of navigational skills and knowledge poses difficulties in situations where navigation systems fail, such as when network connectivity is lost, a battery is drained, or there is not enough time to input a destination into the system. Burnett and Lee (2005) called for a new "learning-oriented" user interface design for navigation systems, and Oliver and Burnett (2008) later observed that adding landmarks and traces of previous trips to the visual map display encouraged cognitive map development. We hypothesize that a learning-oriented navigation system based on dialogue and long-range navigation instructions will be equally or more effective, while providing a more natural interaction modality that does not require visual attention to the navigation interface. In the system we envision, the next instruction offered would be based on a model of the user's navigational knowledge, which would be estimated from observations of past navigation sessions and from dialogue with the user. Such a system may be thought of as an intelligent tutoring system for urban navigation. Following the "scaffolding" paradigm from education research, the strategy of the tutor would be to gradually reduce the level of assistance along frequent routes until the user achieves mastery (Wood et al., 1976). This abstract describes the planned navigation tutor in more detail, as well as results from a preliminary experiment in urban navigation.

The goal of the tutor is to facilitate the user's development of a cognitive map of his or her local area, while also providing assistance for unfamiliar routes. Jackson (1996) found that, for study participants watching a video of a particular route taken from the driver's perspective, the introduction of a narrator reading turn-by-turn instructions caused one group of participants to remember the details of the route less well than another group that had no narrator. Burnett and Lee (2005) made similar observations of users who were asked to complete several routes in a driving simulator, where one group of users was given turn-by-turn guidance as they drove and the other was not as they drove. Based on these results, it appears that drivers are better able to form a cognitive map of an area if assistance is limited. At the same time, if a route is not known to the driver, then he or she will need detailed instructions on at least the first occasion. We propose, therefore, that the tutor will adapt the granularity of the instructions to the user's cognitive map. A fine-grained instruction specifies the next turn only, as in current navigation systems, while a coarse-grained instruction specifies an intersection or landmark that requires multiple turns to reach from the driver's current location.

In order to make appropriate decisions about the next instruction, the tutor will engage the user in dialogue to determine if portions of the user's existing cognitive map are relevant to the route. For example, after the user has input the destination, the system might give the prompt, "Do you know how to get part of the way there?" If the user responds by specifying a landmark he or she knows how to reach, the system might follow up by asking, "Could you tell me how to get there?" The problem then is understanding the user's spoken route instructions, which previous studies have addressed successfully in a limited domain (Johansson et al., 2011; Meena et al., 2012; Meena et al., 2013). Once the user's description of the route is understood and checked for correctness, the system can update its own model of the user's knowledge.

Another important question to consider is the choice of representation for navigational knowledge. We propose to use a modified version of the *conceptual route graph*, in which there is a node for every place where a turn could occur along the route, rather than only at places where turns do occur (Müller et al., 2000; Johansson et al., 2011; Meena et al., 2012; Meena et al., 2013). The rationale for this modification is that the user's knowledge of the segments (triples consisting of an edge, an end node, and a turning action to perform) of one route should transfer to overlapping segments of a different route. Figure 1 shows a part of the representation for two such overlapping routes.

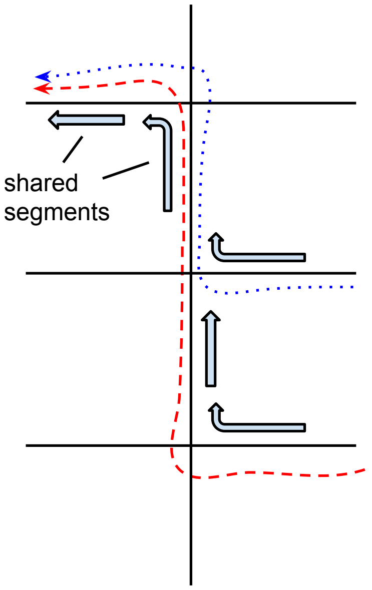


Figure 1: Modified route graph showing two overlapping routes. The routes enter the road drawn vertically from different points, but the overlapping parts of the routes along this road are captured in this representation.

Once the user's knowledge has been assessed, the system must decide on which instruction to offer to the user. Our thought is that the tutor should offer an instruction that matches the user's needs, that is, a turn-by-turn-style instruction when the next part of the route is unfamiliar to the user and a multi-turn instruction when the next part is known. Once an instruction is offered, a brief dialogue between the user and the tutor begins. If the user expresses uncertainty about the instruction offered, the tutor will choose a more fine-grained instruction, which might be comprised of a long-range instruction involving fewer turns or a turn-level instruction. The tutor will need to be mindful of how much time is left before the user approaches the next turn. Similar issues of managing the time available for dialogue during navigation have been considered by Janarthanam et al. (2013). We will consider a reinforcement learning approach to determining an effective strategy for stepping back to more fine-grained instructions in light of the user's response to the initial instruction offered. Reinforcement learning has been been applied to tutorial dialogue systems in other domains to, for example, decide whether to provide feedback after a student's response, or to decide between telling the student a target concept and prompting the student to describe it in his or her own words (Chi et al., 2010; Tetreault and Litman, 2006).

Ahead of more focused studies aimed at addressing the questions raised here, we have carried out a preliminary experiment in a pedestrian navigation scenario. The purpose of the experiment was to make observations about how people give directions remotely. For each of two pairs of participants, one participant acted as the tutor and the other as the user. The tutor directed the user by cell phone to walk to a series of destinations. One tutor could see the user's location in real-time, but the other tutor could not. As a result, the tutor that had access to user location tended to spend more time issuing commands, with limited feedback from the user. By comparison, the tutor that could not see the user's location frequently asked the user to report his location, and the user soon began to volunteer this information as he approached intersections. The user in this latter pair noted that having to report his location helped him to be more aware of where he was. In light of this, we plan to explore prompting the user to self-report progress along a route as another potentially useful tutorial strategy.

We have proposed an approach to designing an intelligent tutoring system for urban navigation. There are several key challenges to be addressed, including how to represent navigational skills and knowledge, how to estimate this information from dialogue with the user and observations of past navigation sessions, and how to choose the next instruction so as to maximize navigational learning over time. If successful, this system would fulfill a need that is not satisfied by current GPS navigation systems: for a navigation system that increases the user's ability to navigate autonomously over time.

References

Gary E Burnett and Kate Lee. 2005. The Effect of Vehicle Navigation Systems on the Formation of Cognitive Maps. In *International Conference of Traffic and Transport Psychology*, pages 407–418.

Min Chi, Kurt VanLehn, and Diane Litman. 2010. Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In *Intelligent Tutoring Systems*, pages 224–234. Springer.

P. G. Jackson. 1996. How Will Route Guidance Information Affect Cognitive Maps? *The Journal of Navigation*, 49(02):178–186.

Srinivasan Janarthanam, Oliver Lemon, Xingkun Liu, Phil Bartie, William Mackaness, and Tiphaine Dalmas. 2013. A Multithreaded Conversational Interface for Pedestrian Navigation and Question Answering. In *SIGDIAL*, pages 151–153, August.

Martin Johansson, Gabriel Skantze, and Joakim Gustafson. 2011. Understanding Route Directions in Human-Robot Dialogue. In *SemDial*, number September, pages 19–27.

R. Meena, G. Skantze, and J. Gustafson. 2012. A Data-driven Approach to Understanding Spoken Route Directions in Human-Robot Dialogue. In *INTERSPEECH*.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2013. Human Evaluation of Conceptual Route Graphs for Interpreting Spoken Route Descriptions. In *CoSLI*, pages 30–35.

R. Müller, T. Röfer, A. Lankenau, A. Musto, K. Stein, and A. Eisenkolb. 2000. Coarse qualitative descriptions in robot navigation. In *Spatial Cognition II*, pages 265–276.

Keith J. Oliver and Gary E. Burnett. 2008. Learning-oriented vehicle navigation systems: a preliminary investigation in a driving simulator. In *MobileHCI*, pages 119–126.

Joel Tetreault and Diane Litman. 2006. Using Reinforcement Learning to Build a Better Model of Dialogue State. In *EACL*.

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.

Improving Knowledge Representation to Speed up the Generation of Grammars for a Multilingual Web Assistant

Marta Gatius
Department of Computer Science, Technical University of Catalonia, Barcelona, Spain
gatius@lsi.upc.edu

Abstract

This paper describes the use of a syntactico-semantic taxonomy to facilitate the generation of grammars for a multilingual web assistant. In particular, it describes the generation of grammars for two different domains: cultural events and medical specialists.

1 Introduction

Most practical conversational systems use semantic grammars adapted to a specific domain because processing results faster and more robust against errors. However, the cost of adapting those grammars to new domains and languages is usually high. To reduce this cost, many systems use semantic models representing domain entities and application specifications to facilitate the generation process. The use of semantic models representing domain concepts is especially appropriate for multilingual systems. Some of those systems use database models (Polifroni et al., 2003; D'haro et al., 2009), others use richer formalisms, such as ontologies (Dzikovska et al., 2003; Cimiano et al., 2007; Sonntag et al.,2007; Nesselrath and Porta, 2011).

In many communication systems only syntax and conceptual levels are distinguished, as in many linguistic works (Jackendoff, 1983). Our approach also distinguishes an intermediate semantic level between these two levels, as proposed in other works (Haliday, 1985; Perkings, 1989; Bateman, 1994).

Our work is on the use of a syntactic- semantic taxonomy to facilitate the generation of grammars in several languages from domain concepts. We have previously used this taxonomy for generating system messages in a dialogue system supporting English, Spanish and Catalan (Gatius et al., 2007). More recently, we have studied its possible usability for a language with a different organization, Hindi, (Gatius and Pailwal, 2013). In this paper, we describe how this taxonomy is used to generate the grammars supporting user's questions on two domains: cultural event and medical specialists.

2 Proposed Knowledge Representation

Our work is focused on the questions about specific domain information the user asks when looking for web information. For this reason, the syntactico-semantic taxonomy we use relates attributes describing domain concepts to the different grammatical structures appearing in questions about those concepts attributes. All the attribute classes distinguished in the taxonomy are necessary to reflect different surface realizations. The basic attribute classes are associated with grammatical roles: participants (**who_does, who_object, what_object**), being (**is**), possession (**has**), descriptions and relationships between two or more objects (**of**) and related processes (**does**). The class **of** is subdivided into three classes: **of_person** representing relations between persons, **of_object** representing relations between objects and **of_description** representing qualities and circumstances related to the concept. The class **of_description** has been subclassified into subclasses representing time, place, manner, cause, quantity, name and type.

Each subclass is associated with several patterns to express questions and answers about the attribute belonging to the class. Additionally, subclasses have been further subclassified if other information relevant for the linguistic realization can be considered, such as having an associated verb or preposition. For example, attributes in the class **of_name** can be realize with general patterns (i.e., *What's <concept-name> name?*), but a new subclass **of_name_person** has been distinguished and it is is associated with the particle title (i.e., *Dr.*).

We have extended theclasses **of_time** and **of_place** by studying the descriptions of time and locations appear in the domains considered. For example, locations of equipments usually consist of a street address or a city zone.

The class **of_time** has been subclassified considering time units and the different forms of expressing them (i.e. weekdays, weekend). Patterns associated with these subclasses cover several forms of expressing time, including, for example, descriptions of intervals of time.

We have used Grammatical Framework (GF) for implementing the grammars because this framework favors the generation of grammars in several languages (Ranta, 2011). In GF, grammars are separated in two parts: **abstract syntax**, defining meaning and **concrete syntax**, mapping meanings to linguistic realization. The abstract syntax is shared across languages while concrete syntax is specific for each language.

In next subsection we describe how we have used the taxonomy to write grammars in GF representing user questions when looking for web information in two domains: cultural events and medical specialists.

2.1 The Generation of Grammars

The process of generating a semantic grammar for a new domain consists of several steps. In a first step, the domain concepts appearing in the communication have to be described by a set of attributes. Then, those attributes have to be classified according to the syntactico-semantic taxonomy. Next, for each language considered, the lexical entries related to the concepts and their attributes have to be incorporated. Using this information, grammars for several languages can be automatically generated. Although the resulting grammars have to be manually supervised and extended, the effort of generating semantic grammars for different languages from scratch is considerably reduced.

Let’s see the process of generating a grammar for the domain of cultural events. There are many web sites giving information on cultural events. Although information appearing in all those web sites is not the same, in most sites there is information about the title, genre, venue and date of the cultural events. For this reason, we have represented this information as the attributes of the concept **Cultural_Event**. Then, those attributes have been classified according to the sintactico-semantic taxonomy, as shown in Figure 1. The attribute **title** represents the name of the event and is obtained at run-time from the web service. It is linked to the class **of_name**. The attribute **genre** has as value the type of the event (i.e., cinema) and is linked to the class **of_type**. The attribute **date** has as value a set of dates and is linked to the class **of_date** (a subclass **of_time**). The attribute **venue** has as value an instance of the concept **Venue** and is linked to the class **of_place**.

Cultural_Event
title : of_name
genre : of_type
venue : of_place
days : of_date

Doctor
name:of_name_person
specialist : of_type
equipment : of_place
days : of_weekdays

Figure 1. Classification of conceptual attributes

Next, the lexical entries associated with the concept (*cultural event* and *take place*) and the attribute values (except those set at run-time and those reused across domains) have to be incorporated. Then, the abstract syntax grammar in GF is obtained. A fragment of the event grammar is shown in Figure 2. As indicated in the header, this grammar uses the grammars **place** and **time**, that define the structures referring to time and locations.

```
abstract event = extends place, time+
flags startcat = askinf
cat askinf; converb;valtype; valplace; valname
fun
geninf : valtype → askinf
conceptualverb : valtype → converb ->valtype
geninfplace : valtype → valplace → askinf
where : valname → askinf
when : valname → askinf
whatname : valtype → askinf
music, cinema, theater, sport, circus: valtype
takes_place : converb
```

Figure 2. A fragment of the abstract grammar

From the abstract grammar, a concrete grammar is automatically generated for each language using the patterns associated with each attribute class.

The process of generating the grammar for the health domain is similar. Figure 1 shows the main domain concept, **Doctor**, and the semantic classification of the attributes describing it.

3 Conclusion

The use of a syntactico-semantic taxonomy acting as an interface between domain conceptual and general linguistic knowledge reduces the effort of generating grammars for new domains and languages. The reuse of grammars defining several forms of expressing time and locations also limits this effort.

References

J.A. Bateman, B. Magnini, and F. Rinaldi, “The Generalized {Italian, German, English} Upper Model,” Proc. European Conference on Artificial Intelligence Workshop on Implemented Ontologies, 1994.

Philipp Cimiano, Peter Haase, and Jörg Heizmann. Porting natural language interfaces between domains: an experimental user study with the ORAKEL system. Proc. 12th International Conference on Intelligent user interfaces, 2007. ACM, New York, NY, USA, 180-189.

L.F. D’Haro, R. Córdoba, J. M. Lucas, R. Barra-Chicote, and R. San-Segundo, “Speeding Up the Design of Dialogue Applications by Using Database Contents and Structure Information,” Proc. SIGdial, 2009, pp. 160-169.

M. Dzikovska, J. F. Allen, and M. Swift, “Finding the balance between generic and domain-specific knowledge: a parser customization strategy,” Proc. Workshop on Knowledge and Reasoning in Practical Dialogue Systems, 2003.

Marta Gatius, Piyush Paliwal. Generating Domain-Restricted Resources for Web Interaction in Several Languages: Hindi, English and Spanish, Proc. The Third International Conference on Advanced Communications and Computation, 2013, Lisbon, Portugal.

Marta Gatius, Meritxell Gonzalez, Elisabet Comelles. A Multilingual Dialogue System for Accessing the Web. Proc. The third International Conference on Web Information Systems and Technologies, 2007, Barcelona, Spain.

M. Halliday. *An introduction to Functional Grammar*. Edward Arnold, London, 1985.

R. Jackendoff. *Semantics and Cognition*. The MIT Press, Cambridge, MA, 1983.

R. Nesselrath and D. Porta, “Rapid Development of Multimodal Applications with Semantic Models,” Proc. IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, 2011, pp. 37-47.

W Perkins. Generation of Natural Language from information in a frame structure. In *Data & Knowledge Engineering*, 4: 101-114, 1989.

J. Polifroni, G. Chung and S. Seneff, “Towars the Automatic Generation of Mixed-Initiave Dialogue Systems from Web Contents,” Proc. EUROSPEECH, 2003, pp. 193-195.

A. Ranta, *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford, 2011.

D. Sonntag, R. Engel, G. Herzog, A. Pfalzgraf, N. Pfleger, M. Romanelli, and N. Reithinger. “SmartWeb handheld – multimodal interaction with ontological knowledge bases and semantic web services,” *AI for Human Computing*, T.S.Huang et al. (Eds.): LNAI 4451, Springer, 2007, pp. 227-295.

Exploring the Role of Laughter in Multiparty Conversation

Emer Gilmartin[1], Francesca Bonin[1,2], Nick Campbell[1], Carl Vogel[2]
Speech Communication Lab[1] / School of Computer Science and Statistics[2]
Trinity College Dublin
`gilmare@tcd.ie`

We report ongoing work on laughter in task-based and social multiparty human conversation, outlining work to date on laughter around topic change, annotation procedures developed and current and future work on laughter in relation to topic change, multimodality, and biosignals.

1 Introduction

Conversation is widely studied through corpus analysis, often concentrating on 'task-based' interactions such as information gap activities (map-tasks [1], spot the difference [2], ranking items [3]) and real or staged business meetings [4], [5]. This task-based dialogue (on which spoken dialogue technology is based [6]) relies heavily on verbal information exchange. However, the immediate task in natural conversation is often not so clear and the purpose of some interaction may be best described as social bonding.

Laughter is universally observed in human interaction. It is multimodal: a stereotyped exhalation from the mouth in conjunction with rhythmic head and body movement [7]. It is part of the gesture call system, older than language [8], predominantly social rather than solo, and aiding social bonding [9]. It punctuates speech [10], and manifests in a range of forms [11]. We investigate laughter *in situ*, using corpora of non-scripted (spontaneous) multiparty interaction: the task-oriented AMI meetings corpus [5], and the conversational TableTalk [12], d64 [13], and DANS corpora. We address laughter and topic change, multimodal aspects of laughter, and the interplay of laughter and bio-signals.

In earlier work on topic change in AMI and TableTalk we found that laughter, and especially shared laughter, is likely near topic change in both corpora, with a stronger effect in TableTalk, and that the number of people laughing together grows with proximity to topic change in TableTalk [14], [15]. These results on multiparty interaction reflect the literature on laughter in two-party dialogue [16], [17], which points towards discourse functions for laughter as a topic termination mechanism. To investigate whether these findings reflect a general phenomenon we extend this temporal analysis to the DANS Corpus. We speculate that laughter may function as a strategy to instigate a topic change, or as a marker of topic exhaustion providing a buffer against an embarrassing silence. We are examining laughter in terms of speaker role (who speaks/laughs first and last, etc.) and turn-taking activity to better understand its function. Our work on multimodality investigates the perception of audio and visual laughter cues by naïve annotators, to investigate whether they can reliably spot unimodally. We are also exploring the interplay of laughter and electro-dermal activity (EDA), linked to levels of emotional arousal [18] and to cognitive load [19]. Social chat has been linked to implicit processing, which is reported to involve lower cognitive load [20], while laughter has been observed to be more frequent in social than in task-based dialogue [21].

2 Annotation of Corpora

The AMI and TableTalk corpora have been annotated previously for laughter. The use of existing annotations is attractive, but some of the annotations exhibited problems outlined in the literature [22], [23], including mixtures of point and interval annotation, laughter annotated on the transcription tier at insufficient granularity – e.g. segmented only to the utterance level rather than to word level, and no method for dealing with laughter when it co-occurs with speech. To address these problems we created a new annotation scheme using Elan [24] with separate laugh tracks for each speaker which we used to re-annotate the TableTalk laughter, using MUMIN [25]. We also noted that some laughs were not sounded, or too quiet to be picked up by microphone. To explore this 'silent' laughter, we expanded our annotation scheme adding two unimodal laughter tiers. In this scheme audio and video laughter is annotated separately - for the video only ('silent') passes, annotators mark laughter intervals on silent video, while the audio only version is created by annotators marking sound recordings of the data. A third annotation is made using both audio and video.

3 DANS Corpus Study

The DANS corpus comprises three sessions of informal English conversation among five participants: two women and three men, four native English speakers and one near-native speaker. The sessions were recorded in a living-room like setting with participants free to speak about any topic. Between two and four participants were on screen at any time. The corpus includes video, audio, and EDA measurements from wrist worn Q-sensors [24]. Laughter annotation was performed as described above. The corpus was also segmented into and turns, topics on the basis of content, and annotated for pauses, gaps, and backchannels. Below we describe preliminary results of our analysis of topic transition relevance and multimodality of laughter in a one-hour section three-party conversation.

4 Results of DANS Corpus study

In the annotations of the three-party one-hour segment of DANS there were 241 laughs of which 49 were solo, with the remaining 192 individual laughs making up 96 shared laughs. There was shared laughter in 92% of topics. The distance from the topic change to the last shared laugh ranged from 10.2 to 0 seconds with 81% of topic changes occurring within 5 seconds of shared laughter.

Table 1 Video (V) and Audio (A) annotation agreement by Participant (P)

P	A and V (%)	A not V (%)	V not A (%)	Agree
1	66(73%)	14 (21%)	11 (16%)	89%
2	51(54%)	2 (2%)	40 (43%)	94%
3	59(83%)	1 (1%)	11 (16%)	93%

We compared silent and sounded laughter annotations in categorical terms; looking at raters' agreement on the incidence rather than the duration of laughter. Table 1 shows the per-speaker (P) results of the laughter annotations. The final column shows the level of annotation by speaker.

We found that most cases where annotations were made on video audio (V not A) involve a combination of head tilting (pitch) and a wide or toothy grin (particularly in Speaker 2). In annotations on the audio but not video (A not V), most involve laughter co-occurring with speech (in Speaker1) with a much smaller number of cases where the annotation was of a short phrase initial or final laugh or snort.

5 Discussion and Conclusions

The results of the topic transition analysis on a section of the DANS corpus are consistent with those obtained in our earlier analysis of AMI and Tabletalk, with a marked preponderance of shared laughter in multiparty social dialogue; this is also in line with reports in the literature on the social nature of laughter. The strong likelihood of laughter before topic change points found in our analysis of DANS echoes the results of our work on TableTalk and AMI, adding further evidence to our claim that laughter is prevalent preceding topic change in social talk. Once completed, our current analysis of participant role around topic change in terms of speaking and laughing will be used to further illuminate the role of laughter around topic change.

The results on multimodality indicate that careful annotation on the audio channel picks up most stereotypical sounded laughter. Humans watching silent video pick up the vast bulk of audio laughter, but can also identify head nods accompanied by a wide grin as laughter. Automatic identification of laughter on the audio stream is possible for stereotypical laughter [27] but requires clean near field audio signals - a limitation for real-world use. Identification on video data is an attractive idea. From our preliminary studies, it appears that humans can identify the incidence of laughter on video alone with high recall but that precision may be an issue. The audio results suggest that a clear distinction needs to be made in our scheme between laughter alone and laughter co-occurring with speech

We have noted the need to re-annotate, and then expand our annotation scheme in view of observations during manual annotation. While data annotation is time-consuming and labour-intensive work, it is invaluable for a fuller understanding of the dynamics of human interaction. Indeed, close examination of data has revealed subtleties that may have been missed had we simply used pre-existing annotations. We have explored laughter in relation to topic change in three different corpora, and have begun to investigate whether laughter can be identified from video or audio alone; a question highly pertinent to social signal processing.

Acknowledgments

This work is supported by the Fastnet Project – Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631, and by the Innovation Academy at Trinity College Dublin.

References

[1] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, and others, 'The HCRC map task corpus', *Lang. Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[2] R. Baker and V. Hazan, 'LUCID: a corpus of spontaneous and read clear speech in British English', in *Proceedings of the DiSS-LPSS Joint Workshop 2010*, 2010.

[3] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, 'From nonverbal cues to perception: personality and social attractiveness', in *Cognitive Behavioural Systems*, Springer, 2012, pp. 60–72.

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, and A. Stolcke, 'The ICSI meeting corpus', in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, vol. 1, pp. I–364.

[5] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos, 'The AMI meeting corpus', in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005, vol. 88.

[6] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, 'An architecture for a generic dialogue shell', *Nat. Lang. Eng.*, vol. 6, no. 3&4, pp. 213–228, 2000.

[7] M. Mehu and R. I. Dunbar, 'Relationship between smiling and laughter in humans (Homo sapiens): Testing the power asymmetry hypothesis', *Folia Primatol. (Basel)*, vol. 79, no. 5, pp. 269–280, 2008.

[8] R. Burling, *The talking ape: How language evolved*, vol. 5. Oxford University Press, USA, 2007.

[9] P. J. Glenn, *Laughter in interaction*. Cambridge University Press Cambridge, 2003.

[10] R. R. Provine, 'Laughter punctuates speech: Linguistic, social and gender contexts of laughter', *Ethology*, vol. 95, no. 4, pp. 291–298, 1993.

[11] J. Trouvain, 'Segmenting phonetic units in laughter', in *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona*, 2003, pp. 2793–2796.

[12] K. Jokinen, 'Gaze and gesture activity in communication', in *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, Springer, 2009, pp. 537–546.

[13] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell, 'D64: A corpus of richly recorded conversational interaction', *J. Multimodal User Interfaces*, pp. 1–10, 2010.

[14] F. Bonin, N. Campbell, and C. Vogel, 'Laughter and topic changes: Temporal distribution and information flow', in *Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 53–58.

[15] E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell, 'Laughter and Topic Transition in Multiparty Conversation', in *Proc. SigDial*, Metz, France, 2013, pp. 304–308.

[16] G. Jefferson, 'A technique for inviting laughter and its subsequent acceptance/declination', *Everyday Lang. Stud. Ethnomethodology*, vol. 79, p. 96, 1979.

[17] E. Holt, 'The last laugh: Shared laughter and topic termination', *J. Pragmat.*, vol. 42, no. 6, pp. 1513–1525, 2010.

[18] M. E. Dawson, A. M. Schell, and D. L. Filion, 'The Electrodermal System', *Handb. Psychophysiol.*, p. 159, 2007.

[19] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, 'Galvanic skin response (GSR) as an index of cognitive load', in *CHI'07 extended abstracts on Human factors in computing systems*, 2007, pp. 2651–2656.

[20] D. Kahneman, *Thinking, Fast and Slow*. Farrar Straus & Giroux, 2011.

[21] R. R. Provine, *Laughter: A scientific investigation.* Penguin Press, 2001.

[22] K. Laskowski and S. Burger, 'Analysis of the occurrence of laughter in meetings.', in *INTERSPEECH*, 2007, pp. 1258–1261.

[23] K. P. Truong and J. Trouvain, 'Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis', *Proc. 4th Int. Worskhop Corpora Res. Emot. Sentim. Soc. Signals*, pp. 20–24, 2012.

[24] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, 'Elan: a professional framework for multimodality research', in *ProcLREC*, 2006, vol. 2006.

[25] J. Allwood, L. Cerrato, L. Dybkær, and P. Paggio, 'The MUMIN multimodal coding scheme', in *Proc. Workshop on Multimodal Corpora and Annotation*, 2004.

[26] M. Z. Poh, N. C. Swenson, and R. W. Picard, 'A wearable sensor for unobtrusive, long-term assessment of electrodermal activity', *Biomed. Eng. IEEE Trans. On*, vol. 57, no. 5, pp. 1243–1252, 2010.

[27] S. Scherer, F. Schwenker, N. Campbell, and G. Palm, 'Multimodal laughter detection in natural discourses', in *Human Centered Robot Systems*, Springer, 2009, pp. 111–120.

Interpreting instructions in a pedestrian routing domain

Jana Götze[1], **Johan Boye**[1], **Morgan Fredriksson**[2], **Jürgen Königsmann**[2]
[1]KTH, School of Computer Science and Communication, 100 44 Stockholm, Sweden
[2]Liquid Media, Hammarby allé 34, 120 61 Stockholm, Sweden
`{jagoetze,jboye}@kth.se` `{morgan,jurgen}@liquid.se`

1 Introduction

Simulated users are important vehicles for testing, development and evaluation of dialogue systems.

We describe an implementation of simulated users that can interact with a dialogue system for pedestrian routing and exploration via written natural language. To emulate real user behaviour, such simulated pedestrians need to have a representation of a user's goals, the past dialogue history, the geographic context, as well as capabilities for generating realistic movement patterns, and for contextually interpretating route instructions.

The dialogue system, described in (Boye et al. 2012) and henceforth called R, uses data from the OpenStreetMap (OSM) geographic database (Haklay, 2008) to construct a route from the user's starting position to his goal, and then give instructions as the user is moving.

A key problem for the simulated user is to interpret such instructions and to resolve the references to objects in the city. Such references form the link between the algebraic and geometric model of the domain, and the communication with the user. For the simulated user it is crucial to correctly interpret instructions like "Turn left onto King's Street.", or questions like "Can you see the statue?" to be able to follow them. In order to interpret questions like the latter, it is also important to have access to information on visibility. In our simulation, visibility is checked on the basis of the OSM database, by continuously calculating whether the line of sight between the user's position and the surrounding objects is intersected by another object such as a building.

2 The pedestrian routing domain

Routing systems have been around quite some time for car navigation, but the pedestrian routing problem is different and in many senses more difficult, as pedestrians have many more options to choose from. Pedestrian routing systems have recently been studied by several researchers (Bartie and Mackaness, 2006; Krug et al. 2003; Janarthanam et al. 2012).

R employs a dialogue strategy of first grounding landmarks with the user, and only then use them in routing instructions. We now want a simulated user that can hold up the user's end of the dialogue to generate dialogues like the following:

1. System: There is a fountain about 35 metres from here. Can you see it?
2. User: Yes.
3. System: Good! Please walk to the left of the fountain.
4. User: (walks)
5. System: Please turn right and walk to the top of the stairs.
6. User: I cannot see any stairs.

In order to generate behaviour that resembles that of a real pedestrian, our simulated user S has a representation of the direction S is currently heading, the desired direction, the set of landmarks currently visible, the landmarks that have been mentioned in previous utterances, and the places that have been visited on previous occasions. S also maintains a representation of the objects in the immediate vicinity in order to generate movement, and to understand relative references like "left" and "right", and a representation of landmarks in its field of vision, but a complete knowledge of the entire city is neither necessary nor desired. The restricted geographic knowledge of the simulated user mimics that of a real pedestrian.

3 Interpretation of utterances

A semantic parser translates natural-language utterances into context-independent expressions in a flat meaning representation language, which is then further processed to resolve references and generate context-dependent interpretations. On the basis of these, goals can be added to the queue of actions for the simulated user to do next.

Here, we consider instructions and propositional questions that require geographical context to find an appropriate referent, as well as utterances that additionally require dialogue context.

For instance, the instruction "Turn left at the junction towards Starbucks on East Crosscauseway", is represented by:

```
dialogAct(inform, X),
X : turn(left, A, B, C),
isA(A, junction),
isA(B, cafe),
isNamed(B, starbucks),
isA(C, street),
isNamed(C,′eastcrosscauseway′)
```

In this expression, the variable `X` is a handle that acts as pointer to the succeeding expression `turn(left, A, B, C)`. The use of handles is inspired by minimal recursion semantics (Copestake et al. 2005). The variables `A, B` and `C` are implicitly lambda-bound, and the purpose of the spatial reference resolution mechanism is to find the identifiers of the nodes that the speaker referred to.

The key semantic predicate for instructions is `turn(Dir, TurningPoint, AimPoint, Street)`

The values of the arguments are constrained by the instruction. The utterance above constrains all four, whereas "Turn left" only constrains the first, and "Go towards Starbucks" only the third.

In order to find concrete nodes to fill in the `TurningPoint` argument, the set of nodes visible from the user's position, and the set of nodes visible from the next goal node is calculated, and a node matching the description is sought among these nodes. The landmarks that `AimPoint` and `Street` refer to are not required to be in view, so the whole set of nearby nodes is searched.

The resolved utterance then becomes:

```
dialogAct(inform, X),
X : turn(left, 21135018, B, 23614881),
isA(21135018, junction),
isA(2156953057, cafe),
isNamed(2156953057,′starbucks′),
isA(23614881, street),
isNamed(23614881,′eastcrosscauseway′)
```

where the lambda-bound variables of the unresolved expression have been substituted with identifiers of nodes and ways. These in turn will be added to the queue of short-term goals. In this example, the user is asked to first go to the junction, and then towards the cafe, i.e. first the junction with ID `21135018`, denoting the `TurningPoint` will be added, then the `AimPoint`.

4 Behaviour generation

The simulated user S generates movement and dialogue behaviour. Dialogue acts that can be expressed are requests for directions ("Directions to Camera Obscura"), requests for instructions ("Where should I go now?"), answers to specific questions ("Yes, I can see Starbucks"), acknowledgements ("Okay"), reports of miscommunication ("I didn't understand that"), reports of success ("Thanks, I can see Camera Obscura"), and a few others.

A dialogue always begins with the simulation stating the long-term goal, e.g. "Directions to Camera Obscura". It then starts walking in a random direction awaiting the first instruction which will lead to one or several short-term goals being put on the goal queue if the instruction is interpreted successfully. If reference resolution does not result in any matching object, a miscommunication report will be generated (e.g. "Go to Starbucks" – "I don't know where Starbucks is."). If S receives no instructions, it will try to guess an appropriate next short-term goal and put it on the queue on its own initiative. Most often, S will continue walking in roughly the same direction as before, but with a small probability it will deviate from its current course and randomly select a new direction.

In addition, the simulated user has a scalar representation of how assertive it is that the current direction is correct. This assertiveness is increased if a given instruction can be interpreted sensibly, e.g. when the instruction is "Turn left", and it is indeed possible to turn left at the next short-term goal. If it is not possible to turn left, assertiveness will be decreased. The assertiveness is also slowly decreased as time elapses without it having received a route instruction, and even more so if S needs to change direction on its own initiative. If the assertiveness value falls below a certain threshold, this information can be used to generate a request for help, e.g. "Where should I go now?".

5 Concluding remarks

We have described an implementation of simulated users in a pedestrian routing domain that can interpret route instructions in their spatial context and dialogue context. Ongoing work includes methods for also modifying the dialogue behaviour using past interactions with real users as well as testing how a simulated user interprets instructions that real users give to describe routes while moving along them, e.g. using the corpus described in (Albore et al. 2013).

References

Albore, A., Boye, J., Fredriksson, M., Götze, J., Gustafson, J. and Königsmann, J. (2013) Final Pedestrian Behaviour Component. Deliverable 2.3.2, SpaceBook project, http://www.spacebook-project.eu/pubs/D2.3.2.pdf.

Boye, J., Fredriksson, M., Götze, J., Gustafson, J. and Königsmann, J. (2012) Walk this way: Spatial grounding for city exploration. *Proc. IWSDS workshop*, Paris, France. Extended version to appear in *Towards a Natural Interaction with Robots, Knowbots and Smartphones*, Springer-Verlag, 2013.

Bartie, P. and Mackaness, W (2006). Development of a speech-based augmented reality system to support exploration of cityscape. *Transactions in GIS*, 10(1):6386.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. Research on Language and Computation, 3(2-3), 281-332.

Haklay, M. (2008) OpenStreetMap: User-generated street maps. *Pervasive computing IEEE*, vol. 7, issue 4, pp. 12–18.

Janarthanam, S., Lemon, O. Liu, X., Bartie, P., Mackaness, W., Dalmas T. and Götze, J. (2012), Integrating location, visibility, and Question-Answering in a spoken dialogue system for pedestrian city exploration. *Proc. SeineDIAL, 16th Workshop on the Semantics and Pragmatics of Dialogue*, Paris, France.

Krug, K., Mountain, D. and Phan, D. (2003) Webpark: Location-based services for mobile users in protected areas. *GeoInformatics*, pp. 26–29.

"That's all. Thank you." Emergence of Formulaic Protocols among Japanese EFL Learners

Yasunari HARADA
Faculty of Law, Waseda University
1-6-1 Nishi-Waseda, Shinjuku-ku,
Tokyo 169-8050, Japan
harada@waseda.jp

Mayumi KAWAMURA
c/o Yasunari Harada
1-6-1 Nishi-Waseda, Shinjuku-ku,
Tokyo 169-8050, Japan
maykmr@gmail.com

Daisuke YOKOMORI
Graduate Schosol of Languages and Cultures, Nagoya University
Nagoya, Aichi 464-8601, Japan
yokomori.d@gmail.com

Tomoko ENDO
Faculty of Humanities and Social Sciences, University of Tsukuba
Tsukuba, Ibaraki 305-8571, Japan
endotomoko@gmail.com

Abstract

This presentation describes how formulaic protocols emerge in oral interactions among Japanese EFL learners in college freshman English classes. In this class, the students are organized into groups of three and respond to questions orally presented by other students. The groupings change each week over a period of 30 weeks in a school year. They are instructed to say their numbers and names each time before they read the questions aloud or answer them. Some choose to do so constantly in Japanese and others in English but some others change the language according to what others do in the same group. How they terminate their responses also change over time and certain relatively fixed patterns of expressions may permeate among the students in a class.

1 Introduction

Acquisition of communicative competence has become one of the most important objectives of English language education in Japan. Asking the right questions at the right time and responding to them directly and immediately are integral parts of successful oral interactions in the North American context, but Japanese learners of English experience linguistic, socio-cultural and cognitive difficulties in doing so. In an effort to remedy this problem, we introduced what we call “oral response practice,” in which students are organized into groups of three and try to respond to questions posed by other students in the same group. Those interactions are recorded with digital audio / video recording equipments and some of the materials are transcribed and annotated.

2 Oral Response Practice

For various reasons, the numbers of the students in the freshman English classes where the data collections take place are from 15 to 36. For each session, ten questions pertaining to one particular topic for the week, such as self-introductions, decisions, and plans for the summer, are prepared in advance and printed on business-card size pieces of paper. The students in one group have three roles to play, the questioner, the respondent and the time-keeper. The questioner picks up one of the ten question cards in turn and reads the question aloud to the respondent twice. The respondent has ten seconds to think and formulate the answer and 45 seconds to speak whatever comes to her/his mind. The time-keeper prompts the respondent by saying “Start!” ten seconds after the question is read for the second time and says “Stop!” 45 seconds later. Then, the three students change their respective roles and go on to the next question.

For digitally recording the interactions, we built a portable audio recording device consisting of one 24-track hard disk recorder, Alesis ADAT HD24 XR, and two 8-channel microphone faders, Alesis MultiMix 12R, with 12 sets of microphone cables and electret-condenser microphones, Sony ECM-360 and started using this in 2005. In addition, since fall of 2006, each time-keeper uses a video camera with a 30GB internal hard-drive, Sony DCR-SR100, together with a wireless Bluetooth microphone, SONY HCM-HW1. In our earlier papers, we described our equipment, procedure and environment for data collection and transcription in more details. (Harada et al., 2008; Huang, C. R. et al., 2010)

3 Phrase-final Vowel Lengthening

The data we collected show a number of interesting interactional phenomena among Japanese EFL learners. Phrase-final Vowel Lengthening is one case. In examples (1) and (2) below, stressed and/or inserted vowels are transcribed with curly brackets and vowel lengthening is marked by colons (:). Underlines mark words with PfVL.

(1) There is{u}:: mountain and{o}:: sea.

(2) … but I{i}: think{u}:: it is more important to have{u}:: fun with friends.

There are some apparently similar phenomena shown by Japanese EFL learners, such as epenthesis, in which the speakers add a vowel after a closed syllable. This is caused partly by the Japanese phonological structure, in which there are basically no closed syllables. Speakers who have learned how to pronounce closed syllables when reading sentences aloud, however, may speak with marked PfVL in their spontaneous speech, which suggests it is caused (at least in part) due to L1 discourse strategy. (Harada et al., 2013)

4 Establishment of Opening Protocols

The students are instructed to say their numbers and names each time before reading the questions aloud twice and responding to the questions. They have the choice of first saying the numbers and then their names or the other way around, and they can do so in English or in Japanese, as in the hypothetical examples in (3) and (4).

(3) 5番、大隈花子です。
Go-ban, OHKUMA Hanako-desu.
five-number Ohkuma Hanako-copula

(4) My number is five and my name is Hanako OHKUMA.

At the beginning of the school year, the students are generally not sure how to say their numbers and names but in a few weeks, most start using one or the other formats. Soon, there would be two types of students, those who establish one format and use it consistently and those who use one or the other and follow what the others in the same group for the week do.

5 Emergence of Ending Protocols

Students do not get any explicit instructions as to how to terminate their responses. At the beginning of the school year, most come to the end of the 45 seconds while they are still trying to think of something to say and there are no specific patterns. As the months go on, though, some start using “Thank you.” or “That’s all.” and those expressions may or may not be employed by other students depending on the class. In one particular class we examined, “That’s all.” gained popularity and propagated among more and more students, as students mixed in different groups each week, as indicated in figure 1 below.

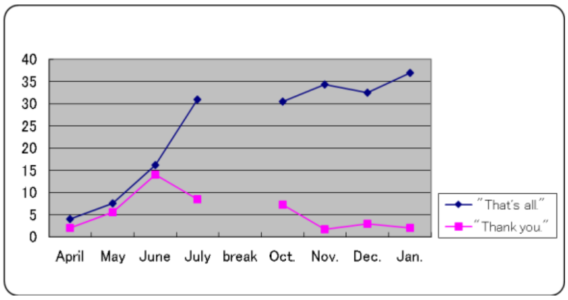


Figure 1 Monthly Total Frequency per Session [April, 2007 through January, 2008]

Acknowledgments

The interactional data reported and discussed here were collected, compiled and annotated in part with financial supports of JSPS/MEXT Grant-in-Aid Scientific Research (B) Project Number 21320109 and Grant-in-Aid Scientific Research (B) Project Number 18320093.

References

Harada, Y., Kawamura, M., Yokomori, D., & Suzuki M. 2013. Data Collection and Annotation of Relatively Spontaneous and Relatively Extended Elicited Utterances by English Learners in Undergraduate Japanese Courses, In Ishikawa, S. (Ed.), *Learner Corpus Studies in Asia and the World: Vol.1 Papers from LCSAW2013*, School of Language & Communication, Kobe University: 179-197.

Harada, Y., Maebo, K., Kawamura, M., Suzuki M., Suzuki, Y., Kusumoto, N., & Maeno, J. 2008. Toward Construction of a Corpus of English Learners’ Utterances Annotated with Speaker Proficiency Profiles: Data Collection and Sample Annotation. In T. Tokunaga and A. Ortega (Eds.), *LKP 2008, Lecture Notes in Artificial Intelligence (LNAI) 4938*. Berlin Heidelberg: Springer-Verlag: 171-178.

Huang, C. R., Cheung, W., Harada, Y., Hong, H., Skoufaki, S. & Chen, H. K. Y. 2010. English Learner Corpus: Global Perspectives with an Asian Focus. In T. Kao & Y. Lin (Eds.), *A New Look at Language Teaching and Testing: English as Subject and Vehicle: Selected Papers from the 2009 LTTC International Conference on English Language Teaching and Testing March 6-7*, Taipei: The Language Training & Testing Center: 85-117.

Aboutness topic, discourse topic and the structure of discourse

Elena Karagjosova
Stuttgart University / Keplerstraße 17
D - 70174 Stuttgart
`elena.karagjosova@ling.uni-stuttgart.de`

The paper addresses the relation between several dimensions along which discourse has been assumed to be structured – topical structure, hierarchical structure, QUD-structure and thematic structure – and points at previously undescribed mismatches between those.

BACKGROUND ASSUMPTIONS: As discourse progresses, the *aboutness topic* of a sentence (Reinhart, 1981; Roberts, 2011; Krifka, 2007) may remain aboutness topic of subsequent sentences in discourse, or the aboutness topic may change (Givon, 1983). This relation between the aboutness topics of subsequent sentences in discourse, which I call the **topical structure of discourse**, constitutes one dimension along which discourse may be structured. Discourse may also be structured along a hierarchy of (explicit or implicit) *questions under discussion* (QUDs) which individual sentences and sequences of sentences in discourse can be seen to answer (von Stutterheim, 1994; van Kuppevelt, 1995; Roberts, 1996). According to (Roberts, 1996), achieving the goal of all discourse, which is the attempt to answer the global QUD ”What is the way things are?”, involves developing sub-goals addressed in terms of answering sub-QUDs. The resulting discourse has a hierarchical **QUD-structure** where each sentence addresses its own, local QUD, and sequences of sentences may answer a joint global QUD. A third dimension is what I call the **thematic structure** of discourse: a sequence of sentences may exhibit certain thematic continuity in terms of a common *discourse topic* (van Dijk, 1976; Asher, 1993; van Kuppevelt, 1995). Finally, discourse may be structured into a hierarchy of discourse units (DUs), where a superordinate sentence/DU may dominate one or several subordinated sentences/DUs. This **hierarchical structure** of discourse is governed by two types of discourse relations between sentences/DUs, *coordinating and subordinating discourse relations* (Asher and Vieu, 2005).[1] The relation between these structures has been discussed before, but I am not aware of a model that takes all these dimensions into consideration. In (van Kuppevelt, 1995; Roberts, 2011), QUD-structure corresponds to thematic structure, since the global QUD of a DU corresponds to the discourse topic of that DU. (Frey, 2005) shows that thematic continuity should be distinguished from topical continuity. Finally, it has been commonly assumed that the QUD-analysis of discourse structure is compatible with the analysis in terms of discourse relations, since the latter can be characterized in terms of implicit questions that relate a sentence to preceding sentences in discourse cf. e.g. (Kehler, 2012).

CLAIMS: I. Thematic structure corresponds to QUD-structure, but there is a mismatch between thematic/QUD-structure and hierarchical structure; II. Topical and thematic structure do not coincide, but thematic structure is sensitive to topical structure, in a way hierarchical structure isn't.

EVIDENCE I: The model of (van Kuppevelt, 1995) which I employ assumes that topicality in terms of a hierarchy of topic-comment structure is a basic organizing principle of discourse structure. A question-based notion of topic-commment structure is used for both individual sentences and larger DUs. A question Q determines a (discourse) topic T defined as a set of possible values (objects, places, times, reasons) of the ”topic term” of the question. One of these values is selected by answer A. The topic term corresponds to *background* in the focus-background distinction (cf. e.g. (Krifka, 2007)), rather than to aboutness topic, and represents a (contextually given or evoked) indeterminacy that needs further specification. The comment C is provided by A. If

---

[1] (Grosz and Sidner, 1986) propose a type of hierarchy governed by *intentions* that is assumed in (van Kuppevelt, 1995) to be related to the QUD-structure and in (Asher and Vieu, 2005) to the subordination/coordination distinction.

the speaker assumes A to be satisfactory for addressee, T is closed off. If not, i.e. if A contains indeterminacies, it triggers a process of subquestioning. Two types of subquestions are distinguished: *quantitative subquestions* asking for additional comment values in case A is incomplete, and *qualitative subquestions*, which either ask (i) for specification of an insufficiently specific value in A and are thus ”goal-satisfying” or (ii) for ”goal-subservient” support (justification, motivation, evidence) of a value in A. Subquestions constitute continuations of the topic constituted by the main question. The model does not consider the hierarchical structure, but it suggests that qualitative subquestions involve elaborations/explanations and thus correspond to these subordinating discourse relations, i.e. when a sentence/DU is elaborating/explaining another sentence/DU, the discourse topic of the superordinate sentence/DU is continued.

This however does not mean that the two structures coincide, and there are cases showing that QUD-structure and hierarchical structure do not fully match. Consider (1) where S_2-S_4 elaborate on S_1. The main QUD of the discourse can be analyzed as (i) *What happened with TC then?* or (ii) *Whom did TC start working for then?* As for option (i), topic T_1 (possible values of ”topic term”) consists of (the set of) things (events) that happened to TC. Answer A_1 specifies working for FFC as one such thing. The possible indeterminacies and subquestions A_1 may trigger involve further values (*What else happened to TC?*; quantitative), elaborations on the value given (*How/Where/When was working for FCC?*; qualitative) or support for the value (*Why did TC work for FFC?*, qualitative). The subquestion that the actual sentence S_2 answers is *How did TC find FFC?*. This is however not an immediate subtopic of T_1 but a subtopic of the subtopic *How was working for FFC?*: the indeterminacy involved in A_1 may be specified by either characterizing the event as a whole (*It was nice/ terrible*) or by characterizing the person involved (*He was nice/terrible*). The actual discourse implements the latter strategy: by specifying how FCC was like, S_2 answers *How was working for FFC?*. This subsubtopic represents an additional level of thematic structure of (1) that is missing in its hierarchical structure: whereas S_2, being an elaboration of S_1, is embedded only one level deep in terms of hierarchical structure, it is embedded two levels deep in terms of thematic structure, hence under T_1, thematic/QUD-structure and hierarchical structure of (1) do not match. (The situation is different in case the main topic is (ii) above: T_1 is ”people TC worked for”, and the indeterminacy is fixed by the immediate (qualitative) subquestion *How was FFC like?*, hence thematic/QUD-structure matches hierarchical structure.)

EVIDENCE II: Van Kuppevelt’s model does not consider topical structure. (Roberts, 2011) suggests that since the aboutness topic of a sentence is part of the (local) QUD the sentence answers, the QUD reflects the aboutness topic. However, this relation between QUD and aboutness topic is sometimes ambiguous. Thus the aboutness topic of S_1 in (2) may be *Hans* or *the concerts*, but in both cases, the QUD may be *How many concerts is Hans giving?*. The relation between thematic and topical structure should be more carefully explored since aboutness topic seems sometimes to play a role in the subquestioning process and thus in determining the thematic structure of discourse: Depending on the combination of aboutness topic and main QUD, the sub-topics that a sentence gives rise to may be different. Thus if the aboutness topic of S_1 in (2) is *Hans* and the QUD *What is Hans doing this week?*, T_1 is ”things that Hans is doing”, and the topic T_2 of the elaborating S_2 should be ”properties of things Hans is doing”. If the aboutness topic is *the concerts*, and the QUD *How many concerts is Hans giving*?, T_1 should be ”number of concerts Hans is giving”, and the indeterminacy that S_2 attempts to resolve is different as it is related to further properties of Hans’ concerts. Topical structure does not seem to play the same role at the hierarchical level as at the thematic level (this being further evidence for the different nature of the two), as suggested by comparing the subordinated structures in (2) and (3). In (3), the discourse referent in the explanation S_2 is not aboutness topic, as shown by the topic test, whereas in (2), the elaborating S_2 introduces new aboutness topic. I.e., forming a subunit does not seem to depend on whether the aboutness topic is continued or not. The contrast between (2) and (3) does not seem to hinge on the type of subordinating relation, as the opposite configuration is also possible, at least in the case of elaboration, cf. (4).

ACKNOWLEDGMENTS: Research reported in the paper is funded by SFB 732, Stuttgart University.

Examples

(1) [Then **Tom Cruise** went to work for F.F. Coppola. [...]]$_{S_1}$ [**Coppola** he found to be ”just like one of the guys.]$_{S_2}$ [And **he** totally trusted me.”]$_{S_3}$ [**He** let me go anywhere I wanted to go with the character”]$_{S_4}$ (from (Roberts, 2011))

(2) [**Hans** will give two concerts this week]$_{S_1}$. [**The first one** will be on Monday in Bochum.]$_{S_2}$ [**The second one** will be on Tuesday in Hamburg.]$_{S_3}$ (from (Frey, 2005))

(3) [**The meeting** is postponed.]$_{S_1}$ [The director is ill.]$_{S_2}$
The meeting is postponed. #About the director, he is ill.

(4) [**Hans** will give two concerts this week]$_{S_1}$. [**He** is playing on Monday in Bochum and on Tuesday in Hamburg.]$_{S_2}$

References

Nicholas Asher and Laure Vieu. 2005. Subordinating and coordinating discourse relations. *Lingua*, (115):591–610.

Nicholas Asher. 1993. *Reference to abstract objects in discourse*. Kluwer, Dordrecht.

Werner Frey. 2005. Pragmatic properties of certain German and English left peripheral constructions. *Linguistics*, (43):89–129.

Talmy Givon. 1983. Topic continuity in discourse. an introduction. In T. Givon, editor, *Topic Continuity in Discourse. A quantitative cross-language study*, pages 1–41. Benjamins, Amsterdam.

Barbara J. Grosz and Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3).

Andrew Kehler. 2012. Cohesion and coherence. In C. Maienborn, K. von Heusinger, and P. Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, pages 1963–1987.

Manfred Krifka. 2007. Basic notions of information structure. In C. Fery and M. Krifka, editors, *Interdisciplinary Studies of Information Structure*, number 6. Potsdam.

Tanja Reinhart. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, (27):53–94.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. In J. H. Yoon and A. Kathol, editors, *OSU Working Papers in Linguistics 49: Papers in Semantics*, Revised version 1998, pages 91–136.

Craige Roberts. 2011. Topic. In Claudia Maienborn Klaus von Heusinger and Paul Porter, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 2. Walter de Gruyter, Berlin.

Teun van Dijk. 1976. Sentence topic and discourse topic.

Jan van Kuppevelt. 1995. Main structure and side structure in discourse. *Linguistics*, (33):809–833.

Christiane von Stutterheim. 1994. Quaestio und Textaufbau. In H.-J. Kornadt, J. Grabowski, and R. Mangold-Allwinn, editors, *Sprache und Kognition*, pages 251–272. Spektrum, Heidelberg.

Analysis and Modeling of Concern Alignment in Consensus-Building

Yasuhiro Katagiri
Future University Hakodate, Japan
katagiri@fun.ac.jp

Katsuya Takanashi
Kyoto University, Japan
takanashi@kyoto-u.ac.jp

Masato Ishizaki
The University of Tokyo, Japan
ishizaki@iii.u-tokyo.ac.jp

Mika Enomoto
Tokyo University of Technology, Japan
menomoto@media.teu.ac.jp

Yasuharu Den
Chiba University, Japan
den@cogsci.l.chiba-u.ac.jp

Shogo Okada
Tokyo Institute of Technology, Japan
okada@ntt.dis.titech.ac.jp

1 Introduction

Dialogue provides a central mechanism with which to negotiate a consensus among ourselves in daily interactions. Consensus can be conceived as a formation of shared commitment on certain choice of future joint actions by a group of people. These actions are often mutually conditional on each other for their successes, and hence, consensus-building has invariably involve some form of management of affective trust relationships between conversational participants. 'Concern Alignment in Conversations' project aims to elucidate this interplay between rational agreement seeking and affective trust management through conversations, based on empirical analyses of real life conversation data and computational modeling of the conversational processes.

2 Concern alignment

Our starting hypothesis is that consensus decision-making processes can conceptually be divided into two parts, concern alignment and joint plan construction, as shown in Figure 1(Katagiri et al., 2011; Katagiri et al., 2012). When a group of people are in a situation 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 then proceed to propose and negotiate on concrete choice of actions (*proposals*) to form a joint action plan. When we decide to go for lunch together, we exchange what each deem relevant in selecting a restaurant, e.g., price, location, cuisine etc., before actually naming individual restaurants. In real life dialogues, these two processes can often be interleaved, people go back and forth between concerns and proposals, and a proposal jointly accepted can produce another set of concerns in implementing it at a finer level of details. Based on this conceptual framework, we have been empirically investigating conversational processes of concern alignment in medical consultation dialogues and exploring to establish a computational model of consensus-building through concern alignment.

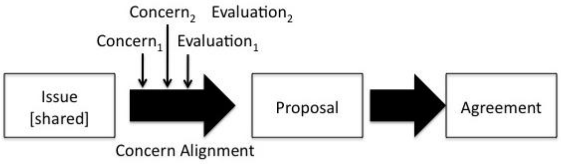


Figure 1: A schematic diagram of the concern alignment process in consensus-building.

3 Corpus-based analysis

Data: We have collected medical counseling dialogues for obese patients. Patients diagnosed as having a metabolic syndrome see expert nurses to get advises on their daily life management. The nurse and the patient discuss and seek a consensus on the methods to improve patient's daily life habits to improve their health. The nurses try to establish affective trust relationships with their patients to keep their patients to stick to their advice after the sessions. We have collected a total of 9 sessions, about 5 hours of dialogues on video. All the sessions were transcribed.

A-B:	C-introduce:(stop smoking)	⇒	C-eval/negative:(no intention)
A-B:	C-introduce:(reduce smoking)	⇒	C-eval/negative:(already tried)
A-B:	C-introduce:(use non-smoking pipe)	⇒	C-eval/negative:(tongue tingling)
B-A:	C-introduce:(cost money)	⇒	C-eval/positive: (acknowledge)
B-A:	C-introduce:(choose tobacco rather than eating)	⇒	C-eval/negative:(not good)
B-A:	C-introduce:(consider when short on money)	⇒	C-eval/positive: (good)
B-A:	C-introduce:(withdrawal syndrome)	⇒	C-eval/positive: (acknowledge)
B-A:	C-introduce:(smoker communication)	⇒	C-eval/positive: (acknowledge)
	⇓		
A-B:	P-introduce: (consider stop smoking when prices go up)		
B-A:	P-accept: (stop smoking when prices go up)		

Figure 2: An example of sequential organization of concern/proposal exchanges in consensus-building dialogue.

Descriptive framework and analysis: Based on the concern alignment ideas, we devised a classificatory scheme for dialogue acts performed by conversational participants in terms of their contribution to concern alignment and joint action plan construction (Katagiri et al., 2013). Figure 2 shows an annotation example of a part of a counseling dialogue session. The analysis captures the process of concern alignment in which the nurse A and the patient B exchange a series of concerns, all related to the patient's smoking behavior, and then focus and settle on a conditional plan for B to stop smoking, based on their responses to raised concerns,

4 Agent modeling for concern alignment

In order to capture and describe the conversational processes of concern alignment in computational terms, we have started to explore agent action selection models using game theoretical ideas.

Incomplete information: A framework for incomplete information games, such as Bayesian games (Harsanyi, 1967), should be employed to capture the process of concern alignment, as agents engaging in negotiation for consensus start with only partial information on their interlocutors' goals and preferences, which is then gradually accumulated through the conversational interactions. Agent types and beliefs about these types include agent utility structures.

Communication game: Instead of treating a conversational exchange as a multi-step extensive-form game, we find it suitable to conceive of a consensus-building session as a communication game (Myerson, 1991), which consists of two phases: communication and action selection. These two phases correspond to concern exchange and proposal exchange steps shown in Figure 1. Based on the information obtained in the communication phase, agents select their actions that maximize the expected utility outcomes.

Concern alignment as update: The process of concern alignment constitutes the presentation and uptake of information on participant utility structures as agent types. Exploration of model behaviors have been underway using schematic interaction settings.

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) 24300061.

References

John C. Harsanyi. 1967. Games with incomplete information played by 'baysian' players, part i. *Management Science*, 14(3):pp.159–182.

Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Mika Enomoto, Yasuharu Den, and Yosuke Matsusaka. 2011. Concern alignment in consensus building conversations. In *the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial2011)*, pages 208–209.

Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Mika Enomoto, Yasuharu Den, and Yosuke Matsusaka. 2012. Negotiation for concern alignment in health counseling dialogues. In *the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial2012)*, pages 173–174.

Yasuhiro Katagiri, Masato Ishizaki, Katsuya Takanashi, Yasuharu, and Mika Emonoto. 2013. An analysis and modeling of mutual trust formation in health guidance dialogues. Technical Report SIG-SLUD-B103: 89-94, Technical Report of Japanese Society of Artificial Intelligence.

Roger B. Myerson. 1991. *Game Theory: Analysis of Conflict*. Harvard University Press.

Towards Structural Natural Language Formalization: Mapping Discourse to Controlled Natural Language

Nicholas H. Kirk
Computer Science Department
Technische Universität München
nicholas.kirk@tum.de

Abstract

The author describes a conceptual study towards mapping grounded natural language discourse representation structures to instances of controlled language statements. This can be achieved via a pipeline of preexisting state of the art technologies, namely natural language syntax to semantic discourse mapping, and a reduction of the latter to controlled language discourse, given a set of previously learnt reduction rules. Concludingly a description on evaluation, potential and limitations for ontology-based reasoning is presented.

1 Motivation

Work towards the formalization of natural language has been pursued on both syntactic and semantic levels. Controlled Natural Languages (CNL) for instance provide an unambiguous set of syntactic rules and a controlled vocabulary (Wyner et al., 2010), while sharing human intelligibility with the original Natural Language (NL) from which it derives (Kuhn, 2013). Approaches to pure semantic formalization have been done via symbolic and distributional characterizations (Blackburn et al., 2001; Harris, 1981), to various extents of compositionality (Clarke, 2012).

An important and structural approach towards formalization of discourse is Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 1993), which makes use of inter- and intra-sentence discourse referents for anaphoric referencing and meaning preservation, and a set of semantic-level constraints over them. DRT maintains transformations to and from logic formalisms (Kamp and Reyle, 1993), and has direct applications within the automated sentence construction domain (Guenthner and Lehmann, 1984; Fuchs et al., 2010). Given the logical and linguistic properties of CNL (e.g. reasoning, paraphrasability, human- and machine- readability) the author stresses that a successful mapping between NL and CNL can enable language based cognition of simple autonomous software assistants, for reasoning and as interface to both peers and humans.

2 Concept

Given such rationale, the community should formulate a methodology for operating a reduction of sentence-level natural language discourse, to a discourse representation formulated in a target controlled natural language.

The author presents a possible pipeline abstraction of preexisting state-of-the-art means, as described in Figure 1. In particular, source channel text normalization (C1) to regularize erroneous phonetic transcriptions and spelling; a text to grounded Discourse Representation Structures (DRS) parser (C2) which works thanks to Combinatory Categorial Grammar (CCG), i.e. a grammar formalism that allows a computationally efficient interface between syntax and structural semantics (Curran et al., 2007). The implemented form has already achieved optimal results and can produce Discourse Representation Structures as output (Bos, 2008); a previously trained sentence-level Support Vector Machine (SVM) rule classifier, which identifies the types of NL to CNL reductions that should be operated (C3). A similarly implemented classifier is present in literature (Naughton et al., 2010). We then have a syntactic manipulation engine to transform the natural language input DRS into a set of compliant CNL DRS instances (C4), subject to the previously obtained classification results. Such classification (C3) should account for, for instance:

- intrinsically ambiguous natural language

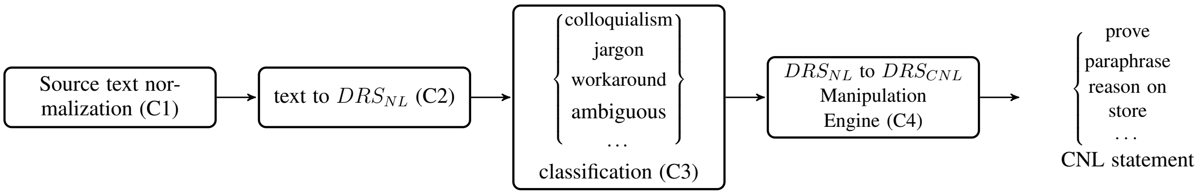


Figure 1: Representation of an abstract structure-level only NL to CNL manipulator

syntactic constructs

- ambiguous anaphoric reference resolution
- conscious constraining decisions on the expressiveness of specific CNL constructs

The full enumeration of reduction case reasons is application domain-dependent and require an aprioristic study that can be performed online and in a supervised manner, for instance with active learning techniques. A possible target CNL which has proven robustness and reliability is ACE (Fuchs et al., 2006), which has DRS to CNL verbalization functionalities, as well as paraphrasing, proving and inference reasoning capabilities. Figure 2 shows a simple instance of the presented pipeline, which requires manipulation via sostitution of the unigram ”linguistics” with the trigram ”a linguistic class”.

NL: "Harris can teach linguistics on Tuesdays."

⇓

ACE: "Harris can teach a linguistic class on Tuesday."

Figure 2: Example of an NL sentence instance and a possible semantic-preserving reduction to ACE

Evaluation Evaluation should mainly assess, via the use of human evaluation, if given an arbitrary sentence related to the application domain, the meaning of this has been successfully conveyed to the target controlled sentence. For instance, a threshold of satisfactory quality in action-oriented tasking domains (Nyga and Beetz, 2012) can be if arguments of intra-, mono-, ditransitive verb arguments have been preserved, together with correct anaphoric resolution. Evaluation will also assess domain-specific classification rates and computational efficiency.

Limitations The presented architecture does not make assumptions on the content of the predicates that are represented by words, given that the manipulation is operated only at a structural level, i.e. within the boundaries of DRS expressiveness. For a deeper predicate-related alignment, further considerations regarding lexicon should be made, to provide word sense and Part-Of-Speech (POS) mappings between source vocabulary and target controlled vocabulary.

Potential Current statistic-based web search approaches that make use of word n-gram models can exploit a more structural, discourse oriented approach. Formalization enables logic satisfiability check of manipulated NL questions via reduction and reasoning on First Order Logic (FOL) clauses. The expressiveness of the latter would also allow reasoning as Constraint Satisfaction Problems (CSP), i.e. a widely adopted mathematical formalism that expresses real-world decision problems as unary and binary constraints over finite variable domains. To pursue the example in Figure 2, admitting other ontological knowledge of lecturers' availability and ability, we could formulate an NL question (that becomes a formal ACE question) to ask for solutions to a simple timetable scheduling CSP problem, where the domains are the possible lecture days and types, and the constraints are the required lecture types and time precedence relations between them.

3 Future Work and Conclusions

This concept-only presentation hopes to have briefly highlighted the potential that such abstract CNL-based architecture can have, above all within the context of artificial assistants, as a means of interface, logic and combinatorial problem reasoning in ontology-based applications. If compliant with CNL rules, a specific set of syntactically reduced NL statements can seamlessly interface humans and machines while maintaining intelligibility and logical properties, such as entailment verification and inference. Future work should focus on implementation and efficiency verification of the stated architecture, to then investigate predicate-level (lexical) semantic alignment, to step towards (quasi-) complete sentence-level natural language formalization.

References

Patrick Blackburn, Johan Bos, Michael Kohlhase, and Hans De Nivelle. 2001. Inference and computational semantics. In *Computing Meaning*, pages 11–28. Springer.

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

Daoud Clarke. 2012. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

James R Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 33–36. Association for Computational Linguistics.

Norbert E. Fuchs, Kaarel Kaljurand, and Gerold Schneider. 2006. Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces. In *FLAIRS 2006*.

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2010. Discourse Representation Structures for ACE 6.6. Technical Report ifi-2010.0010, Department of Informatics, University of Zurich, Zurich, Switzerland.

Franz Guenthner and Hubert Lehmann. 1984. Automatic construction of discourse representation structures. In *Proceedings of the 10th international conference on Computational linguistics*, pages 398–401. Association for Computational Linguistics.

Zellig S Harris. 1981. *Distributional structure*. Springer.

Hans Kamp and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Number 42. Springer.

Hans Kamp. 1981. A theory of truth and semantic representation. *Formal semantics-the essential readings*, pages 189–222.

Tobias Kuhn. 2013. The understandability of owl statements in controlled english. *Semantic Web*, 4(1):101–115.

Martina Naughton, Nicola Stokes, and Joe Carthy. 2010. Sentence-level event classification in unstructured texts. *Information retrieval*, 13(2):132–156.

Daniel Nyga and Michael Beetz. 2012. Everything robots always wanted to know about housework (but were afraid to ask). In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, October, 7–12.

Adam Wyner, Krasimir Angelov, Guntis Barzdins, Danica Damljanovic, Brian Davis, Norbert Fuchs, Stefan Hoefler, Ken Jones, Kaarel Kaljurand, Tobias Kuhn, et al. 2010. On controlled natural languages: Properties and prospects. In *Controlled Natural Language*, pages 281–289. Springer.

Abduction and parameterised semantic composition in speech-gesture integration

Udo Klein, Hannes Rieser, Florian Hahn and Insa Lawler
Collaborative Research Center
“Alignment in Communication” (CRC 673)
Bielefeld University, Germany

Abstract

An important feature of speech-gesture integration is that speech and gesture content influence each other in interpretation. To analyse this, we sketch an approach based on three assumptions: (i) the interpreter infers by abduction an explanation for why a particular gesture is synchronized with a particular utterance (part), (ii) semantic composition amounts to the identification of free variables (called parameters), and (iii) abductive inferences determine which parameters to identify during the semantic composition of speech and gesture content.

1 Introduction

Face-to-face communication is often accompanied by gestures: Speakers point at things or shape their contours. Foundational questions arise: What is a gesture’s meaning and how is it determined? And, given that speech and gesture meaning interact, how can they be fused? The issue of speech-gesture integration (SGI) has been studied in various paradigms such as Montague Grammar, HPSG and theories of Discourse and Dialogue; it is also the focus here. We will demonstrate a methodology for integrating verb phrases with accompanying gestures based on parameterised semantic composition.

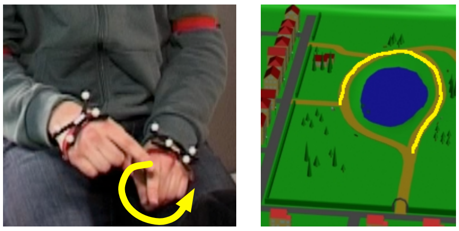


Figure 1: Gesture (left) depicting the agent’s trajectory around a pond (right)

Our work is based on data from a systematically annotated corpus, the Bielefeld-Speech-and-Gesture-Alignment-corpus (SaGA; (Lücking et al., 2013)), which consists of 25 dialogues of dyads engaged in route descriptions. Consider the following example (cf. Fig. 1). The speaker in Fig. 1 describes how to walk through a park passing a pond. While uttering *Gehst quasi drei Viertel um den Teich herum* (Engl.: ‘(You) roughly walk three quarters around the pond (around)’), a round shape is depicted in overlap with the expression *drei Viertel um den Teich herum*. Importantly, this expression specifies that the agent’s trajectory is three quarters around the pond, but it does not specify the actual shape of the trajectory (the shape of the pond does not necessarily determine the shape of the agent’s trajectory around the pond). As a result of being synchronized with this expression, the gesture can be interpreted as specifying that the shape of the agent’s trajectory around the pond is circular. To analyse this, we propose (i) that the interpreter infers by abduction an explanation for why the gesture is synchronized with this utterance part, (ii) semantic composition amounts to the identification of free variables (called parameters), and (iii) the abductive inference enriches semantic composition of speech and gesture content by determining which parameters to identify. In our example the inferred explanation for the synchronicity of gesture and utterance is that the finger trajectory approximates the shape of the agent’s trajectory around the pond.

2 Motivating parameterised semantics for SGI

Previously, we have developed a general methodology for SGI, abstracting from speech acts. We worked out a λ-calculus based solution in which speech meaning is type-lifted to a function which takes gesture meaning as an argument and yields the integrated meaning (cf. Röpke et al., 2013). Here, we present an alternative approach in order to explicitly model the way in which abductive inferences enrich the semantic composition of speech and gesture content (cf. Hobbs (2008) for an overview of abduction in natural language understanding). We propose that the basic principle of semantic composition is conjunction (cf. Pietroski, 2005) relative to (i) a coordination scheme (cf. Fine, 2007) which specifies which free variables in the conjuncts are to be identified, and (ii) a systematic renaming of the remaining free variables in order to avoid accidental identification (cf. Kracht, 2013). To illustrate, the composition of the two formulas $P(x4, x5)$ and $Q(x5, x6)$ relative to the coordination scheme $\{\langle x4, x6\rangle\}$ results in the formula $P(x40, x50) \wedge Q(x51, x61) \wedge x40 = x61$. The free variables in the left and right conjuncts have been suffixed by a 0 and 1, respectively, in order to avoid the accidental identification of the two $x5$ occurrences. The coordination scheme indicates which free variables get identified. An important consequence of using parameterised semantic composition in SGI is that speech and gesture content can be used to instantiate rules used in abductive inferences, and thus to determine the parameter(s) of the utterance content that the gesture content specifies. Moreover, speech and gesture content can be combined without having to change the combinatory potential (and thus the logical type) of utterance content.

3 Analysis

Applying this theory to our example, the composition of $\llbracket$ *drei Viertel* $\rrbracket$ and $\llbracket$ *um den Teich herum* $\rrbracket$ conjoins the two formulas and identifies the degree parameters by adding the equation $d0 = d1$:

$$\frac{drei\ Viertel}{\underline{d} = 0.75} \bullet_{\{\langle \underline{d}, \underline{\underline{d}}\rangle\}} \frac{um\ \ldots\ herum}{\begin{array}{l}\mathsf{mover}(e) = x \wedge \\ \mathsf{trajectory}(x, e) = t \wedge \\ \mathsf{around}(t, r, d) \wedge \\ r = \iota x.\mathsf{pond}(x) \wedge \\ \underline{\underline{d}} \geq 0.5\end{array}} =$$

$$\frac{drei\ Viertel\ \ldots\ herum}{\begin{array}{l}d0 = 0.75 \wedge \\ \\ \mathsf{mover}(e1) = x1 \wedge \\ \mathsf{trajectory}(x1, e1) = t1 \wedge \\ \mathsf{around}(t1, r1, d1) \wedge \\ r1 = \iota x.\mathsf{pond}(x) \wedge \\ d1 \geq 0.5 \wedge \\ \\ \underline{d0} = \underline{\underline{d1}}\end{array}}$$

($x1$ is the entity moving in $e1$, $x1$'s trajectory in $e1$ is $t1$, $t1$ circumscribes some pond $r1$ to a degree $d1 \geq 0.5$, and $d1 = 0.75$.)

The semantic integration of speech and gesture is based on an abductive inference involving the following gesture interpretation rule (GIR):

GIR If parameter p of gesture content $\llbracket G \rrbracket$ approximates some parameter p' of utterance content $\llbracket U \rrbracket$, then G is synchronized with U.

The most plausible instantiation of **GIR** in the utterance context is that the parameter g of $\llbracket G_1 \rrbracket$ representing the finger trajectory approximates the parameter t of $\llbracket U_1 \rrbracket$ representing the mover's trajectory around the pond. Since G_1 is synchronized with U_1, the interpreter can infer by abduction that indeed the parameter g of $\llbracket G_1 \rrbracket$ (the finger trajectory) approximates the parameter t of $\llbracket U_1 \rrbracket$ representing the mover's trajectory around the pond. This inference enriches the semantic composition of gesture and utterance by adding the formula $\mathsf{approx}(g, t)$ to the gesture content, and by specifying the coordination scheme for composition, namely that the trajectory $\underline{t}$ which g approximates is to be identified with the trajectory $\underline{\underline{t}}$ of the mover x in e:

$$\frac{G}{\begin{array}{l}\mathsf{circular.traj}(g) \wedge \\ \mathsf{approx}(g, \underline{t})\end{array}} \bullet_{\{\langle \underline{t}, \underline{\underline{t}}\rangle\}} \frac{drei\ \ldots\ herum}{\begin{array}{l}\mathsf{mover}(e) = x \wedge \\ \mathsf{traj}(x, e) = \underline{\underline{t}} \wedge \\ \mathsf{around}(t, r, d) \wedge \\ r = \iota x.\mathsf{pond}(x) \wedge \\ d = 0.75\end{array}} =$$

$$\frac{G + drei\ \ldots\ herum}{\begin{array}{l}\mathsf{circular.traj}(g0) \wedge \\ \mathsf{approx}(g0, t0) \wedge \\ \\ \mathsf{mover}(e1) = x1 \wedge \\ \mathsf{traj}(x1, e1) = t1 \wedge \\ \mathsf{around}(t1, r1, d1) \wedge \\ r1 = \iota x.\mathsf{pond}(x) \wedge \\ d1 = 0.75 \wedge \\ \\ \underline{t0} = \underline{\underline{t1}}\end{array}}$$

The resulting multimodal representation thus expresses that the mover's trajectory around the pond is a circular one.

To conclude, we propose a novel approach to speech-gesture integration, in which the gesture interpretation is determined by context-dependent abductive inferences and gets integrated with the utterance denotation by parameterised semantic composition. In future work, we intend to compare this approach with our λ-calculus based approach, focusing in particular on how these approaches explain the fact that speech and gesture interpretation mutually influence each other.

Acknowledgments

This research is supported by the German Research Foundation (DFG) in the Collaborative Research Center 673 “Alignment in Communication”.

References

Fine, K. (2007). Semantic Relationism. Blackwell, Oxford.

Hobbs, J. R. (2008). Abduction in Natural Language Understanding. In Horn, L. and Ward, G., editors, The Handbook of Pragmatics, pages 724–741. Blackwell Publishing Ltd.

Kracht, M. (2013). Agreement Morphology, Argument Structure and Syntax. Unpublished manuscript.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2013). Data-based Analysis of Speech and Gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its Applications. Journal on Multimodal User Interfaces, Vol. 7(1-2), pages 5–18.

Pietroski, P. (2005). Events and Semantic Architecture. Oxford University Press.

Röpke, I., Hahn, F., and Rieser, H. (2013). Interface constructions for gestures accompanying verb phrases. In Proceedings of 35th Annual Conference of the German Linguistic Society (DGfS), pages 295–296.

Using Learned Predictions of User Utterances to Decrease Distraction

Fredrik Kronlid
Talkamatic AB
Första Långgatan 18
413 28 Göteborg
Sweden
`fredrik@talkamatic.se`

Staffan Larsson
Talkamatic AB
Första Långgatan 18
413 28 Göteborg
Sweden
`sl@talkamatic.se`

Pontus Wärnestål
Halmstad University
Box 823
301 18 Halmstad
Sweden
`pontus.warnestal@hh.se`

Abstract

Driver distraction is one of the most common causes of accidents. By having a dialogue manager request predicted user answers from a user model instead of asking the user, we can reduce the number of utterances in the dialogue and thereby reduce the time that the user is distracted.

1 Background

1.1 Driver Distraction

As interaction complexity in the car increases due to more advanced infotainment systems, and peripheral technologies in the form of smartphones and tablets, drivers are often executing several tasks in parallel to the primary task of driving. The increased functionality of car information and entertainment systems has resulted in large hierarchical information architectures that prolong interaction time and that may thereby negatively affect safety as well as user experience (Dagmar & Albrecht, 2009). According to the 100-Car Study (Neale et al., 2002), non-primary task distraction is the largest cause of driver inattention.

The goal of the work described in this paper is to design an in-vehicle information system that enables shorter and more efficient interaction in the form of natural language dialogues. The basic assumption is that using apps and services in an in-vehicle context inherently leads to distraction, and that reducing interaction time will reduce driver distraction.

1.2 TDM

Based on Larsson (2002) and later work, Talkamatic AB has developed the Talkamatic Dialogue Manager (TDM). TDM provides a general interaction model based on patterns found in human-human dialogue, resulting in a high degree of naturalness and flexibility which increases usability. TDM offers integrated multi-modality which allows user to freely switch between modalities. The model is domain-independent which means that dialogue behaviour can be altered without touching application properties and vice versa.

1.3 Grounding in TDM

Grounding is, roughly, the process of making sure that dialogue participants agree on what has been said so far and what it meant. TDM has an extensive model of grounding (Larsson, 2002). It operates on different levels: *Perception*, *Semantic Understanding*, *Pragmatic Understanding* and *Acceptance*. System feedback (positive, negative and in some cases interrogative) can be generated on each level. Examples: “I didn’t hear” – negative perception; “Madonna, is that right?” – interrogative semantic understanding; “OK” – positive acceptance.

2 Learning and Classification

Many dialogue applications require the user to answer a number of questions. To make dialogue shorter, we have extended TDM so that it tries to predict user answers on the basis of a user model learned from observations of user behaviour. As an illustration, we use a road information application which tries to predict the user’s destination and thereby eliminate the need to ask the user about this.

2.1 Selection of learning Method

Initially, more complex learning methods (MDP, POMDP) were explored, but the KNN (K-Nearest Neighbours) were considered the best method. An important advantage is that KNN can learn from a relatively small set of observations. This is in contrast to the MDP and POMDP methods, which require large amounts of data to generate useful behaviour. A potential drawback of KNN is that this model cannot model sequences of user behaviours.

On the basis of user studies, it was decided that the most important user model parameters was position, day of the week and hour of the day. The training data were simulated and correspond to the behaviour of an archetypical persona provided by the user partner in the project.

The learning part of the system listens for a number of events, such as “start-car”, “stop-car” etc.. From these events and information about current position, the time of the day and the day of the week, the system creates new data instances. The system thus learns how the user’s destination varies depending on these parameters. When the dialogue manager requests a prediction of the destination, the KNN algorithm tries to find the K data points closest to the present data point, and the top alternatives are returned to the dialogue manager together with confidence scores indicating the reliability of the predictions.

3 Integration of Classifications into TDM

3.1 Grounding uncertain information

We treat the information emanating from the user model as uncertain information about a (predicted) user utterance. Hence, the same mechanisms used for grounding utterances have been adapted for integrating user model data.

3.2 Integrating Classifier Output

TDM is based on the Information State Update (ISU) approach to dialogue management. The rule for integrating the user model data is a standard ISU rule, consisting of preconditions and effects on the information state. The information state in TDM is based on that of the system described in Larsson (2002).

If the user model data is sufficiently reliable to be trusted, the ISU rule described informally below triggers:

Preconditions If the user is the latest speaker and if there is a propositional answer from the user model resolving a question in the current plan, and if the confidence score reported from the user model is above a certain level, the rule should be applied.

Adaptation Effects Applying the rule means that we should accept the propositional answer (include it into the shared commitments), and – depending on the confidence score – give feedback to the user by enqueuing an appropriate feedback move on the agenda. We isolate three different cases when it comes to the feedback:

- For highly probable answers, we embed the feedback move into the next system utterance, e.g. “Which route do you want to take to work?”. The user can always reject the predictionby requesting another destination.
- For relatively certain answers, the feedback move (positive understanding) can be realised as “I assume you’re going to work”. If the user says “no”, the answer is rejected, but silence is interpreted as acceptance.
- For uncertain answers the feedback would be “To work, is that correct?” (interrogative understanding). In this case, the user needs to explicitly accept the proposed answer. Otherwise, the user is prompted for an answer.

3.3 GUI Behaviour

If the ISU rule above does not apply because of too low confidence scores, user model information is still used in the GUI. When a Wh-question is raised by the system, the GUI always presents a list of possible alternatives. High-confidence alternatives are highlighted and sorted before the other alternatives in the list.

4 Conclusions and further work

We have designed and implemented a mechanism which, when exposed to repeated patterns of use, simplifies and shortens the dialogue. It remains for future work to establish that this actually reduces the distraction rate of drivers. We also want to test the performance of the learning mechanism by training it on real observations of user behaviour (as opposed to simulated data).

The current mechanism only predicts answers to individual system questions, which may result in suboptimal behaviour in cases where there are dependencies between the questions pertaining to some task. An interesting area for future work is to instead predict *sequences* of answers; however, this would require a more powerful learning and classification mechanisms.

Acknowledgements

This work was carried out within the FFI project “Safe Speech by Knowledge” (2012-00941), funded by VINNOVA, Volvo Car Corporation and Talkamatic.

References

Dagmar, K. and S. Albrecht *Design space for driver-based automotive user interfaces*. In Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications. 2009, ACM: Essen, Germany. 2009.

Staffan Larsson. *Issue-Based Dialogue Management*. PhD thesis, University of Gothenburg, 2002.

Staffan Larsson, Alexander Berman, and Jessica Villing. *Adding a speech cursor to a multimodal dialogue system*. In INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy. 2011. pages 3319-3320.

Vicki L. Neale, Thomas A. Dingus, Sheila G. Klauer, Jeremy Sudweeks, and Michael Goodman. *An overview of the 100-car naturalistic study and findings*. 2005.

Sharon L. Oviatt, Rachel Coulston, and Rebecca Lunsford. *When do we interact multi- modally?: cognitive load and multimodal commu- nication patterns*. In ICMI , pages 129-136. 2004.

Making Human-Robot Quiz Dialogue More Conversational by Adding Non-Quiz Talk

Ivana Kruijff-Korbayová and **Bernd Kiefer**
DFKI, Saarbrücken, Germany
{ivana.kruijff,bernd.kiefer}@dfki.de

Ilaria Baroni and **Mattia Coti Zelati**
Fondazione Centro San Raffaele, Milan, Italy
{baroni.ilaria,cotizelati.mattia}@hsr.it

Abstract

We present on-going work from the EU-funded project Aliz-E on long-term social human-robot interaction. Our conversational system implemented on the Nao robot engages a user in several activities chosen to support children hospitalized due to diabetes. Here we focus on Quiz, a knowledge-exchange activity about health-related concepts. We recently started to add non-activity talk to Quiz, with the aim to encourage the child to disclose its habits and experiences related to nutrition and diabetes. We will present initial observations about the structure of non-activity talk and the responses elicited from children in an experiment.

Children are keen users of new technologies and new technologies can provide interesting opportunities to enrich children's experience, e.g., for educational and therapeutic purposes (Tartaro and Cassell, 2006). As children are not small adults, it is necessary to research their specific needs and develop systems that address them. The project ALIZ-E develops cognitive robots for adaptive social interaction with young users over several sessions in real-world settings (Belpaeme. et al., 2013).[1] The conversational system developed in ALIZ-E using the Nao robot[2] engages a user in several different activities chosen with regard to the target application domain of the system, namely long-term interaction with children hospitalized due to metabolic disorders, in particular diabetes. More detail about the ALIZ-E system are available in (Kruijff-Korbayová et al., 2011; Kruijff-Korbayová et al., 2012). Here we concentrate on Quiz, a knowledge-exchange activity meant to support learning of health-related concepts.

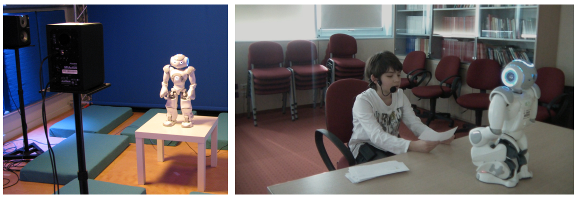


Figure 1: *Left*: Nao in a measurement setup in a sound lab. *Right*: The Quiz activity during experiments in the San Raffaele hospital in Milan.

During the Quiz activity the child and the robot ask each other series of multiple-choice questions from various domains, including diabetes and healthy nutrition, as well as sport, geography and history. Besides activity-specific conversation, the interactions involve also a social component, such as greetings and introductions. During an activity the robot provides performance feedback to the user. The social aspect here requires careful handling of the evaluation process so as not to discourage the user with negative feedback. As the system is designed to have multiple encounters with a user, the robot's behavior differs in various aspects from the first session (meeting for the first time) to the subsequent sessions ("knowing" the user and their performance). To increase the feeling of familiarity between the robot and the child, the robot uses the child's name and it refers to experiences in previous sessions.

Due to its predominantly verbal character and naturally constrained interaction structure the Quiz activity is a good testbed for speech-processing technologies.

Recently we started to experiment with adding what we call *non-activity talk* to the Quiz interactions. We conceive of non-activity talk as being similar in character to *small talk*. However, small talk is typically considered to be "a conver-

---

[1] The EU-FP7 project ALIZ-E (ICT-248116): `http://aliz-e.org/`

[2] `http://www.aldebaran-robotics.com/en`

sation for it's own sake", "an informal type of discourse that does not cover any functional topics or any transactions that need to be addressed"[3], while our non-activity talk has specific topics and a defined purpose. Its purpose is to elicit talk from the child, in particular, to encourage it to disclose its habits and experiences related to nutrition and diabetes. If successful, non-activity talk could provide a therapeutically valuable instrument.

In collaboration with two psychologists at the San Raffaele hospital in Milan we defined the following topics for non-activity talk:

- Hobbies: typical day; activities in spare time
- Diabetes: checking glycemia; checking insulin; injections; hypoglycemia
- Nutrition: eating habits; food choices
- Friends: discussions about diabetes; handling diabetes when with friends
- Adults: behavior w.r.t. diabetes; advice

We then formulated system utterances eliciting talk about these topics (several utterances per topic). For the time being these utterances are implemented as canned text in the system. The system might for example say:

- Hobbies: *What do you like to do in your spare time?*
- Diabetes: *At home, do you check glycemia yourself?* or *If your glycemia is low, what do you do?*
- Nutrition: *How often do you eat fruits and vegetables?*
- Friends: *When you go out with your friends, do you bring with you glucometer and insulin?*
- Adults: *How do your parents behave with you with respect to diabetes?*

At relevant points during the Quiz, such as a question with semantically related content, the robot tries to engage the child in non-activity talk. It first says something to "escape" from the Quiz talk, e.g., *Now, I am curious about something.* Then it raises the respective topic as illustrated above. The utterances on a given topic can be chained in order to create a more complex extended sub-dialogue. The system resumes the Quiz activity by saying, e.g., *OK, now let's do another quiz question.*

We carried out a Wizard-of-Oz experiment with children at a Diabetes Summer Camp in Italy in August 2013, where we collected first insights about non-activity sub-dialogues in sessions with 14 different children. In the system used in the experiment the Wizard simulated the recognition and interpretation of the user's speech and the next system action w.r.t. the non-activity talk. The next system action in the Quiz activity was selected and verbalized automatically, while the Wizard had the possibility to override the automatic selection if needed. Spoken output was synthesized using Mary TTS (Schröder and Trouvain, 2003) with an italian voice developed in the project (Kruijff-Korbayová et al., 2012).

In the poster we will present the overall scenario and experiment setup and then focus on our initial observations about the structure of non-activity talk in the collected dialogues, the responses elicited from the children and how the non-activity talk influenced the dialogue flow.

References

T. Belpaeme., P. Baxter, R. Read, Wood., H. Cuayáhuitl, B. Kiefer, S. Racioppa, Kruijff-Korbayová, G. I., Athanasopoulos, V. Enescu, R. Looije, M.A. Neerincx, Y. Demiris, R. Ros-Espinoza, A. Beck, L. Cañamero, A. Hiolle, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, , and R. Humbert. 2013. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):35 – 53.

I. Kruijff-Korbayová, G. Athanasopoulos, A. Beck, P. Cosi, H. Cuayáhuitl, T. Dekens, V. Enescu, A. Hiolle, B. Kiefer, H. Sahli, M. Schröder, G. Sommavilla, F. Tesser, and W. Verhelst. 2011. An event-based conversational system for the nao robot. In *IWSDS 2011*, Granada, Spain, Sep.

I. Kruijff-Korbayová, H. Cuayáhuitl, B. Kiefer, M. Schröder, P. Cosi, G. Paci, G. Sommavilla, F. Tesser, H. Sahli, G. Athanasopoulos, W. Wang, V. Enescu, and W. Verhelst. 2012. Spoken language processing in a conversational system for child-robot interaction. In *Workshop on Child-Computer Interaction*.

M. Schröder and J. Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

A. Tartaro and J. Cassell. 2006. Using virtual peer technology as an intervention for children with autism. In *Universal Usability: Designing Computer Interfaces for Diverse User Populations*, pages 231–262. New York, John Wiley & Sons, Ltd.

---

[3] `http://en.wikipedia.org/wiki/Small_talk`

Visual distraction test setup for an multimodal in-vehicle dialogue system

Staffan Larsson
Sebastian Berlin
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden
`sl@ling.gu.se`
`sebastian.berlin@gu.se`

Anders Eliasson
Mecel AB
Box 140 44
SE-400 20 Gothenburg
Sweden
`anders.eliasson@mecel.se`

Fredrik Kronlid
Talkamatic AB
Första långgatan 18
SE-413 28 Gothenburg
Sweden
`fredrik@talkamatic.se`

Abstract

The goal of the SIMSI (Safe In-vehicle Multimodal Speech Interaction) project is threefold. Firstly, to integrate a dialogue system for menu-based dialogue with a GUI-driven in-vehicle infotainment system. Secondly, to further improve the integrated system with respect to driver distraction, thus making the system safer to use while driving. Thirdly, to verify that the resulting system decreases visual distraction and cognitive load during interaction. This demo paper describes the test environment designed to enable evaluation of the system, and the planned visual distraction tests.

1 Background

1.1 Driver distraction and safety

Driver distraction is one common cause of accidents, and is often caused by the driver interacting with technologies such as mobile phones, media players or navigation systems. The so-called 100-car study (Neale et al., 2005) revealed that secondary task distraction is the largest cause of driver inattention, and that the handling of wireless devices is the most common secondary task. The goal of SIMSI is to design systems which enable safe interaction with technologies in vehicles, by reducing the cognitive load imposed by the interaction and minimizing head-down time.

1.2 The SIMSI Dialogue System

Based on Larsson (2002) and later work, Talkamatic AB has developed the Talkamatic Dialogue Manager (TDM) with the goal of being the most competent and usable dialogue manager on the market, both from the perspective of the user and from the perspective of the HMI developer. TDM provides a general interaction model founded in human interaction patterns, resulting in a high degree of naturalness and flexibility which increases usability. Also, TDM reduces complexity for developers and users, helping them to reach their goals faster and at a lower cost.

TDM supports multi-modal interaction where voice output and input (VUI) is combined with a traditional menu-based GUI with graphical output and haptic input. In cases where a GUI already exists, TDM can replace the GUI-internal interaction engine, thus adding speech while keeping the original GUI design. All system output is realized both verbally and graphically, and the user can switch freely between uni-modal (voice or screen/keys) and multi-modal interaction.

To facilitate the browsing of lists (a well known interaction problem for dialogue systems), Talkamatic has developed its Voice-Cursor technology[1] (Larsson et al., 2011). It allows a user to browse a list in a multi-modal dialogue system without looking at a screen and without being exposed to large chunks of readout information. A crucial property of TDM's integrated multimodality is the fact that it enables the driver of a vehicle to carry out all interactions without ever looking at the screen, either by speaking to the system, by providing haptic input, or by combining the two. We are not aware of any current multimodal in-vehicle dialogue system offering this capability.

While TDM offers full menu-based multimodal interaction, the GUI itself is fairly basic and does not match the state of the art when it comes to graphical design. By contrast, Mecel Populus is an commercial-grade HMI (Human Machine Interface) with professionally designed visual output. We have previously produced an integration of the TDM and Mecel Populus platforms(Larsson et al., 2013) to establish a commercial-grade HMI for experiments and demonstrations.

---

[1] Patent Pending

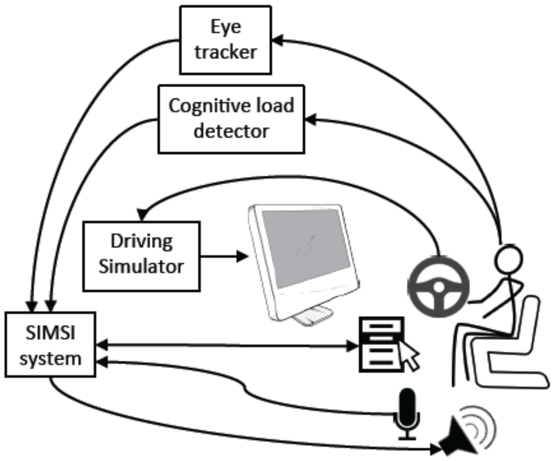


Figure 1: SIMSI test environment overview

2 Test environment

One goal of SIMSI is to conduct ecologically valid test of the applications, and to let the results of these tests feed back into the development of the system. Basically, we want to find the best interaction solutions and to verify these experimentally, especially in cases where it is not intuitively clear what is best. This involves implementing variants of a behaviour, testing them on naive users, collecting data from these interactions, and establishing statistically significant results based on the collected data.

The test environment consists of two parts, apart from the dialogue system: a driving simulator (SCANeR from Oktal) and an eye tracker (Smart Eye Pro from Smarteye). In later tests we will also include instruments for measuring cognitive load.

3 Visual distraction tests

The main point of the visual distraction tests is to investigate how the "eyes-on-road" time during interaction varies between different modality conditions. The eyetracker equipment will be used for capturing where the driver is looking. In addition, driving behaviour (including lane deviation) and dialogue state (including task success) is continously logged.

The following three conditions will be tested:

- GUI only (haptic only in, graphics only out)
- Multimodal with voice cursor (haptics and speech in, graphichs and speech out)
- GUI with voice cursor (haptics only in, graphichs and speech out)

Resources permitting, we may also test two additional conditions:

- multimodal without voice cursor (haptics and speech in, graphichs and speech out)
- speech-only with voice cursor (haptics and speech in, speech only out)

For each condition, we will be using two difficulty levels: easy and difficult. For both levels, the task is to drive along a softly curving road while keeping distance to one car in front of you and one car behind you. In the easy condition, the other cars have a constant speed. In the difficult condition, the other cars are speeding up and braking erratically, and the car behind you may indicate (by honking its horn) that you're going too slow.

This way of testing, which we informally refer to as the "annoying cars" setup, differs from existing experimental setups such as the ConTRe task (Engonopoulos et al., 2008). In the latter, the driver tries to match two vertical lines representing the vehicles position and the target (reference) position. Our setup has the advantage of being more realistic, although we acknowledge that it is still far from driving in real traffic. (On the negative side, our setup does require a full driving simulator environment, which the ConTRe task does not). Initial tests will be carried out to verify the adequacy if the "annoying cars" setup for our purpouses.

The application used in the tests has very basic phone functionality: browsing a list of contacts, and calling people up. At regular intervals, the driver receives a spoken instruction (with a voice different from the dialogue system), e.g. "You just remembered you need to call up Ashley on her mobile number.". The user should then carry out this instruction as efficiently and completely as possible.

We hypothesise that in the GUI only condition, there will be less eyes-on-road time than in the other two conditions, since the driver does not have to look at the screen in order to complete the task. Apart from testing this hypothesis, we are generally interested in which condition(s) gives the best results with respect to eyes-on-road time, task success, task completion time and usability (rated subjectively using a questionnaire).

We will demonstrate the SIMSI system, the three test conditions, and parts of the test environment.

References

Nikolaos Engonopoulos, Asad Sayeed, and V Demberg. 2008. Language and cognitive load in a dual task environment. In *CogSci 2013*, pages 2249–2254.

Staffan Larsson, Alexander Berman, and Jessica Villing. 2011. Adding a speech cursor to a multimodal dialogue system. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, 2011*, pages 3319–3320.

Staffan Larsson, Sebastian Berlin, Anders Eliasson, and Fredrik Kronlid. 2013. Integration and test environment for an in-vehicle dialogue system in the SIMSI project. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 2002–2004.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University.

Vicki L. Neale, Thomas A. Dingus, Sheila G. Klauer, Jeremy Sudweeks, and Michael Goodman. 2005. An overview of the 100-car naturalistic study and findings. Technical report.

A Corpus-Based Study of *Accept* and *Assessment* in SWBD-DAMSL

Yanjiao Li
Department of Chinese,
Translation and Linguistics
City University of Hong Kong
Hong Kong SAR
`yanjiaoli2-c`
`@my.cityu.edu.hk`

Alex C. Fang
Department of Chinese,
Translation and Linguistics
City University of Hong Kong
Hong Kong SAR
`acfang@cityu.edu.hk`

Jing Cao
School of Foreign Languages
Zhongnan University of
Economics and Law
Wuhan, P. R. China
`cecilia_cao`
`@znufe.edu.cn`

Abstract

A series of studies have focused on backwards functions in conversation, especially to explore the distinction among *backchannel/acknowledgement*, *accept* and *yes-answer,* while for *accept* and *assessment/appreciation,* little attention has been obtained. This paper describes a quantitative investigation into two dialogue acts *accept* and *assessment/appreciation*, exhibiting their similarities and differences in the preceding contexts and lexical realization, which is expected to help the automatic detection of dialogue acts.

1 Introduction

Backwards communicative functions (Jurafsky et al., 1997), as one class of dialogue acts (DAs) in conversation, serve to give feedback to the interlocutor, playing a significant role in the interpretation of language in interaction. Data-intensive studies have been conducted in English language to detect discourse structure for speech recognition and understanding tasks (Jurafsky et al., 1998) as well as the design of spoken dialogue system (Bunt, 2012; Gravano et al., 2012). Particularly, it has been widely noted that *backchannel/acknowledgement*, *accept* and *yes-answer* strongly overlap in lexical realization (e.g. Jurafsky et al., 1998; Shriberg et al. 1998; Gardner, 2001; Stolcke et al. 2000; Gravano et al., 2007). While it has been discovered in the current study that *accept* ("aa") and *assessment/appreciation* ("ba") also share similarities, which has not been discussed in past studies. For example, the same utterance string has been found to be tagged as "aa" and "ba" (e.g. "that's right", "that's true" and so on) with similar preceding tags. The current study presents qualitative evidence as the first step for a broad analysis of various backwards functions, to show similarities and differences in the preceding contexts and lexical realization, which we believe is crucial to the successful automatic detection of DAs.

2 Corpus Resource

This study uses Switchboard Dialogue Act corpus (www.ldc.upenn.edu), which comprises 1,155 transcribed telephone conversations, totaling in 223,606 utterances (Fang et al., 2011). In this corpus, the segmented unit for utterances is defined as "slash-unit" (Meeter et al., 1995: 16), which has been coded with DA information according to the SWBD-DAMSL coding scheme (Jurafsky et al., 1997). In this scheme, "aa" refers to the case where the speaker explicitly accepts a proposal, or makes agreements with previous opinions (Jurafsky et al., 1997: 37), while "ba" is defined as "a backchannel/continuer which functions to express slightly more emotional involvement and support than just 'uh-huh'" (P48). See Table 1 for the basic statistics of the two DAs in the corpus.

	Total		Intersection			
	Tokens	Types	Types		Token	
aa	10,136	1,374	139	10.1%	7,718	76.1%
ba	4,523	1,621		8.6%	1,077	23.8%

Table 1. Basic statistics of "aa" and "ba"

139 utterances in the intersection indicate they can function as "aa" or "ba" in the corpus, accounting for 76.1% and 23.8% respectively in terms of tokens. It demonstrates that 76% of utterances in "aa" reoccur as "ba", implying significant lexical similarities between them.

3 Empirical Statistics

DA from the previous utterance as one of the predictors helps to improve the accuracy for recognition of some DAs (Coria & Pineda, 2007). So investigation here is to explore whether the preceding DA tags can be used to differentiate “aa” and “ba”. Since “aa” and “ba” are both positive responses to what has been stated by others rather than by themselves, their previous contexts are restricted to immediately previous utterances uttered by others, tags of which have been partly listed in Table 2.

Pre-aa	F	%	Pre-ba	F	%
sv	4204	41.5	sd	2497	55.3
sd	1907	18.8	+	569	12.6
+	1378	13.6	sv	400	8.8
%	556	5.5	%	309	6.8
bf	451	4.4	b	134	3.0
ba	248	2.4	ny	95	2.1
^2	160	1.6	sd^e	81	1.8
ad	153	1.5	aa	49	1.1
b	126	1.2	x	49	1.1
qh	91	0.9	sd(^q)	40	0.9

Table 2. Top ten previous DA tags

As can be noted in Table 2, the top four tags of previous contexts for both “aa” and “ba” include “sv”, “sd”, “+” and “%”, and constitute about 80% in both cases, indicating “aa” and “ba” share lots of similarities in immediately preceding DA tags. Nevertheless, “sd” and “sv” exhibit their own preference: one is more likely to appear in the preceding of “ba”, while the other prefers “aa”. For 139 utterances lying in the intersection, it is expected to check whether their preceding tags could offer more cues to disambiguation. Table 3 presents the top ten tags of preceding contexts for these 139 utterances.

Pre-aa	F	%	Pre-ba	F	%
sv	3153	40.9	sd	495	46.0
sd	1511	19.6	sv	205	19.0
+	1079	14.0	+	171	15.9
%	407	5.3	%	70	6.5
bf	357	4.6	b	23	2.1
ba	193	2.5	sd^e	17	1.6
^2	141	1.8	ny	11	1.0
ad	106	1.4	x	10	0.9
qh	72	0.9	ba	9	0.8
b	63	0.8	sd(^q)	8	0.7

Table 3. Top ten previous DA tags of the intersection

Similarly, the top four tags account for 80% preceding contexts, which is in line with those in Table 2. Therefore, it can be inferred “aa” and “ba” occur in overlapping environments. As for lexical realization, normally it is believed that “aa” and “ba” are totally different, but their intersection manifests in some cases one utterance can function as “aa” or “ba”. Table 4 exhibits these utterances as well as the results of significant test.

Utterances	F-aa	F-ba	Log-likelihood	Sig.	
yeah	2993	6	2136.15	0	***
right	948	6	640.87	2.2E-141	***
yes	565	2	395.05	6.6E-88	***
no	445	3	299.41	4.4E-67	***
that’s great	1	88	196.73	1.1E-44	***
…	…	…	…	…	…
I’m sure <laughter>	2	1	0.0085	0.9265	
exac-	2	1	0.0085	0.9265	
<laughter> true	2	1	0.0085	0.9265	
that’s right <laughter>	16	7	0.0019	0.9652	

Table 4. Intersection between “aa” and “ba”

Log-likelihood and significant values, calculated by log-likelihood ratio calculator (Xu, 2009), are used to compare two models, expressing “how many times more likely the data are under one model than the other” [1]. The larger the log-likelihood value is, the smaller the significant value is, so the difference between the two sets is more salient. Cases indicated by the symbol “*” are significant, more likely performing one function over the other. Statistically, lexical realization for 29% (40/139) utterances in the intersection can be the predictor to distinguish “aa” from “ba”. However, cases like “*that’s right*” show little preference.

4 Conclusion

This paper presents a corpus-based investigation into “aa” and “ba” in Switchboard Dialogue Act Corpus. According to a batch of quantitative evidence and analyses, “aa” and “ba” share similar contexts expressed and lexical realization. Also, they exhibit a few statistical differences, which can be used to differentiate them, providing evidence to the automatic detection of DAs. In the future, a broader study of various backwards communicative functions will be further conducted. Apart from the preceding contexts and lexical realization, a more specific view will be held on grammatical and syntactic constructions that have been overlooked before.

---

[1] http://en.wikipedia.org/wiki/Likelihood-ratio_test

References

Bunt, H. 2012. *The Semantics of Feedback*. In 16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012), pp 118–127.

Coria, S. R., & Pineda, L. A. 2007. *Prediction of Dialogue Acts on the Basis of the Previous Act.* Procesamiento del Lenguaje Natural, 39.

Fang, A., H. Bunt, J. Cao, and X. Liu. 2011. *Relating the semantics of dialogue acts to linguistic properties: A Machine Learning Perspective through Lexical Cues.* In Proceedings 5th IEEE International Conference on Semantic Computing, Stanford University, Palo Alto.

Gardner, R. 2001. *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins Publishing.

Gravano, A., Hirschberg, J., Benus, S. 2012. *Affirmative Cue Words in Task-oriented Dialogue.* Computational Linguistics, 38(1):1–39.

Gravano, Agustin, Stefan Benus, Hector Chavez, Julia Hirschberg, and Lauren Wilcox. 2007. *On the Role of Context and Prosody in the Interpretation of "okay".* In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), pp 800-807, Prague.

Jurafsky, D., Shriberg, E. and Biasca, D. 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13.* University of Colorado, Boulder Institute of Cognitive Science Technical Report 97-02.

Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. 1998. *Lexical, Prosodic, and Syntactic Cues for Dialog Acts.* In Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers, pp. 114-120.

M. Meeter and A. Taylor. 1995. *Dysfluency Annotation Stylebook for the Switchboard Corpus*. available online at ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.

Shriberg, E., R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema.1998. *Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?* Language and Speech, 41(3-4): 443-492.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. 2000. *Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech.* Computational Linguistics, 26(3): 339-373.

Xu Jiajin. 2009. *Log-likelihood Ratio Calculator.* Retrieved from http://ishare.iask.sina.com.cn/f/13589753.html. Beijing: National Research Centre for Foreign Language Education.

Multimodal Propositions?
From Semiotic to Semantic Considerations in the Case of Gestural Deictics

1 Introduction: Gestural Deixis

We call utterances that comprise elements that are perceived by different sense modalities *and* are coded according to a non-linguistic code *multimodal utterances* (cf. (Fricke, 2012)). An example for non-linguistically coded signs are *indexicals* like pointing gestures, which, according to Peirce, bear some *nomological*, *causal* or *attentional* relation to their objects (CP 1.372, 2.248, 2.285 [1]) In situated dialogue, interlocutor's frequently use multimodal utterances like definite descriptions plus pointing gestures in an exophoric way. Accordingly, such deictic acts are a starting point for looking for multimodal propositions: Whereas the linguistic expressions are treated as arbitrary symbols that are interpreted intensionally with respect to a world or a circumstance according to some standard model theory, the nonlinguistic element, if it indeed follows a nonlinguistic code, must, by definition, interpreted in a different way. At least three distinctions have been made at first, however (Levinson, 2008):

- In the most direct way, viz. *gestural deixis*, the pointing identifies a concrete, perceptible object (or event, property, etc., depending on your metaphysical stance).
- The point or region or object in space pointed at in an utterance situation can stand as a proxy for some spatial configuration or referent in the described situation (*transposed deixis*).
- In *symbolic deixis*, the indicated thing is used as a case of *deferred reference* (Nunberg, 1993), that is, standing for something which is somehow related to the indicated entity.

The different uses of deictic gestures are well documented in the literature on co-verbal gestures (see (Fricke, 2007), (Lascarides and Stone, 2009), (Alahverdzhieva and Lascarides, 2011)). They necessitate a distinction between something pointed at in an utterance situation and something referred to in the described situation. We take up the terminology of (Kühnlein, 1999) and call the former *index* and the latter *referent*. Accordingly, the task for a semantics of speech and co-verbal pointing gestures has two aspects:

1. provide an account for the index;
2. provide information for resolving the referent (maybe in pragmatics).

The main focus here is on the first aspect.

2 Significance of Pointing Gestures

Putting theoretical as well as empirical insights together (e.g., (Reimer, 1991), (Bangerter and Oppenheimer, 2006), (Rieser, 2004)), the resulting picture on gestural deictics like *This N* plus pointing is as follows (cf. Figure 1):

1. The demonstrative *This* is an attentional index according to the Peircean distinction motivated above, which shifts the attention of the addressee towards the gesture.
2. The gesture in turn projects a "search space" in terms of a spatial cone extension ((Kranstedt et al., 2006)). The gesture is a causal index, since it is directly affected by the location of the intended referent (cf. the respective remarks in Section 1).
3. The nominal expression *N* finally picks out the referent from the search space by virtue of descriptive conventions.

In order to capture the spatial nature of gestural deictics, we employ a situation semantics-related model with a rich spatial structure, resting on the central notion of oriented vector spaces. An oriented vector space relates to a pointing gesture in the following, twofold way:

---

[1] We follow the convention to quote the *Collected Papers of Charles Sanders Peirce* (Peirce, 1965) in terms of volume (v) and paragraph (p): 'CP v.p'.

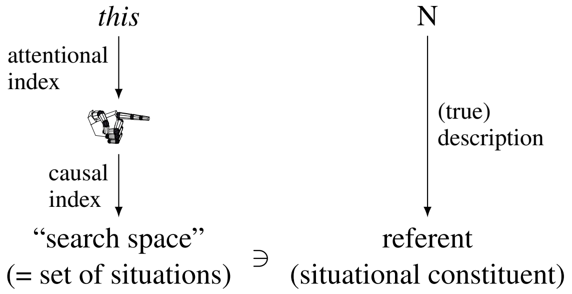

this
N
attentional index
causal index
(true) description
"search space" (= set of situations)
∋
referent (situational constituent)

Figure 1: Collaboration in gestural deictics of the form *This N.*

1. The origin of the speaker's vector space **V** in the current utterance situation (i.e., space(speaker(s))) provides the Bühlerean (Bühler, 1999) Origo for pointing gestures;
2. The pointing cone is projected into the direction determined by the orientation of the index finger with reference to the orienting axes.

Having introduced an oriented vector space according to the two steps above, the spatial extension of a pointing gesture can be specified in terms of sets of vectors. Suppose a pointing gestures G ahead, straight away from the speaker's body. Then the region that G encircles is the set of vectors emanating in the origin of **V** in the direction of **FT**. The corresponding cone covers the following subspace $r(G) \in R \subset \mathbf{V}$: $r(G) = \{\mathbf{v} \,|\, \text{proj}_{\text{LONG}}\,\mathbf{v} < \text{proj}_{\text{VERT}}\,\mathbf{v} \wedge \text{proj}_{\text{LONG}}\,\mathbf{v} < \text{proj}_{\text{LAT}}\,\mathbf{v}\}$ (where $\text{proj}_y\,\mathbf{u}$ is the orthogonal projection from vector **u** onto line y). The subspace defined this way is quite large so that further constraints for instance in terms of angular specifications should be given. However, angular modification does not affect the account sketched here in principal.

The situational extension of a pointing gesture can then be specified in terms of the set of situations which regions that have relatively maximal intersections with $r(G)$:

(1) $[\![\text{☞}]\!]^{\mathscr{M}} = \{e \,|\, \text{region}(e) \cap r(G) \rightsquigarrow \max_i\}.$

'$\rightsquigarrow \max_i$' picks out the i situations that have the largest overlap with the pointing cone. This, the function produces an ordering on situations pointed at, decreasing according to their intersection area with the pointing cone. That is, the spatial extension is assumed to be parameterized. Of course, the best guess at first is to choose the situation with maximal intersection (what corresponds to setting parameter $i = 1$). However, any $i < 2$ does no harm as long the maximal situation provides only on entity that fits the nominal description. This includes plural entities in case of plural nominals. The parameterized treatment leaves a great way of modeling freedom for taking semantic-pragmatic interface issues into account, but is out of scope here. In particular, the spatial model with anchored and oriented vector spaces and the form-based, perspectival interpretation of pointing gestures spells out the cone stipulations verbalized in (Lascarides and Stone, 2009, p. 44). It also gives an account for "referents at certain coordinates" as assumed by (Alahverdzhieva and Lascarides, 2011, p. 17).

Semiotics	*Semantics*
affectedness	form-based interpretation
non-symbolic code	perspectivity
focusing attention	reflexivity

Table 1: Contraposing semiotic and semantic features of gestural deictics.

3 Discussion

In which ways, if any, are propositions for gestural deixis multimodal?

1. Multimodal propositions are *reflexive*: situation s occurs both as the described entity and as part of the description (cf. Figure 1).
2. The interpretation of a gesture G, by means of the determination of $r(G)$, is essentially *affected by the form* of the gesture.
3. The demonstration part of gestural deictics is *perspectival* by depending on the speaker's orientation in space.

The comparison of semiotic theorizing and semantic modeling is summarized in Table 1; it shall not be claimed, however, that pairs of cells are related in a one-to-one manner.

This model not only captures a great deal of semiotic and empirical insights briefly introduced above, it also goes beyond the formal analyses proposed *in this respect* so far.

References

Katya Alahverdzhieva and Alex Lascarides. 2011. An HPSG approach to synchronous speech and deixis. In Stefan Müller, editor, *Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar*, pages 6–24.

Adrian Bangerter and Daniel M. Oppenheimer. 2006. Accuracy in detecting referents of pointing gestures unaccompanied by language. *Gesture*, 6(1):85–102.

Karl Bühler. 1999. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. UTB, Lucius & Lucius, Stuttgart.

Ellen Fricke. 2007. *Origo, Geste und Raum*. Linguistik – Impulse & Tendenzen. Walter de Gruyter, Berlin and New York.

Ellen Fricke. 2012. *Grammatik multimodal. Wie Wörter und Gesten zusammenwirken*, volume 40 of *Linguistik – Impulse und Tendenzen*. De Gruyter, Berlin and Boston.

Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006. Deictic object reference in task-oriented dialogue. In Gert Rickheit and Ipke Wachsmuth, editors, *Situated Communication*, pages 155–207. Mouton de Gruyter, Berlin.

Peter Kühnlein. 1999. Dynamics of complex information. Proc. of ESSLLI. Utrecht, Netherlands.

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

Stephen C. Levinson. 2008. Deixis. In *The Handbook of Pragmatics*, chapter 5, pages 97–121. Blackwell.

Geoffrey Nunberg. 1993. Indexicality and deixis. *Linguistics and Philosophy*, 16(1):1–43.

Charles Sanders Peirce. 1965. Collected papers of Charles Sanders Peirce. repr. from 1932.

Marga Reimer. 1991. Do demonstrations have semantic significance? *Analysis*, 51(4):177–183.

Hannes Rieser. 2004. Pointing in dialogue. In *Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue*, Catalog '04, pages 93–100, Barcelona.

Establishing a communication system: Miscommunication drives abstraction

Gregory Mills
Institute for Language, Cognition and Computation
University of Edinburgh
gmills@staffmail.ed.ac.uk

One of the central findings in research on the emergence of communication systems is that interlocutors rapidly converge on a shared set of contracted referring expressions (Krauss and Weinheimer, 1966; Clark, 1996) which become progressively systematized and abstract. This occurs for a wide range of referents, e.g. when referring to spatial locations (Garrod and Doherty, 1994), music (Healey et al, 2002), concepts (Schwartz, 1995), confidence (Fusaroli et al., 2012), and temporal sequences (Mills, 2011). Systematization of referring expressions occurs across modalities – in spoken interaction (Pickering and Garrod, 2004), text-based interaction (Healey and Mills, 2006) and in graphical, mediated interaction (Healey et al., 2007). This pattern is observed both when interlocutors are faced with the task of describing unfamiliar referents (Galantucci, 2005), as well as when interlocutors already possess referring expressions suitable for individuating the referents (Pickering and Garrod, 2004). Even when referring expressions are given experimentally, interlocutors coordinate on the semantics of their referring schemas (Larsson, 2007). Further, the quality of the interaction directly affects the development of coordination. If interlocutors are prevented from providing each other with feedback, e.g. by being prevented from drawing on each other's drawings, this impedes the development of systematicity (Healey, 2007).

Cumulatively, these findings suggest that interaction in dialogue places important constraints on the semantics of referring expressions. However, there is currently no consensus about how best to account for how coordination develops, e.g. whether it occurs as a natural consequence of exposure to another's linguistic output (Kirby, Cornish, Smith, 2008), as a consequence of mutual priming (Pickering and Garrod, 2004), or via interlocutors providing each other with positive evidence of understanding (Clark, 1996).

To investigate in closer detail the development of referential coordination, we report a variant of the “maze task” (Pickering and Garrod, 2004). Participants communicate with each other via an experimental chat tool (Mills and Healey, 2006), which interferes with the unfolding dialogue by inserting artificial probe clarification requests that appear, to participants as if they originate from each other. The clarification requests signal apparent miscommunication of participants’ referring expressions.

Participants who received clarification requests performed better at the task, and also converged more rapidly on more abstract and more systematized referring expressions. We demonstrate how this beneficial effect is due to the artificial clarification requests amplifying naturally occurring miscommunication: When interlocutors establish a novel communication system, signals of miscommunication provide interlocutors with evidence of negative understanding of each other’s referring expressions. Consequently, amplifying these signals yields enhanced problem detection and improves recovery from error. We argue that these results show that abstraction and systematicity of communication systems is driven by negative evidence: miscommunication drives convergence.

References

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

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. Cognitive Science: 29(5), 737.

Garrod, S. (1999). The Challenge of Dialogue for Theories of Language Processing. Language Processing. Hove: Psychology Press.

Garrod, S and Doherty, D. (1994). Conversation, co-ordination and convention:"Cognition, 53, 181–215.

Healey, P. G. T. (1997). Expertise or expert-ese: The emergence of task-oriented sub-languages. Proceedings of CogSci (pp. 301–306).

Healey, P. G. T., Swoboda, N., Umata, I., & King, J. (2007). Graphical Language Games. Cognitive Science, 31, 285–309.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: PNAS. 105(31), 10681–10686.

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation and the encoding of referents in verbal communication. JPS 4, 343–346.

Larsson, S. (2007). Coordinating on ad hoc semantic systems in dialogue. In Proceedings of DECALOG.

Mills, G J. (2011). The emergence of procedural conventions in dialogue. In Proceedings of CogSci, Boston.

Pickering, M. J., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. Behavioural and Brain Sciences, 27(2), 169–190.

Schwartz, D. L. (1995). The emergence of Abstract Representations in Dyad Problem Solving. Journal of the Learning Sciences, 4(4), 321–354.

Pragmatic Markers of Common Ground in Tertiary Institution Students' Sexual Discourse in Lagos State, Nigeria

Oni-Buraimoh Olawunmi PhD
Department of English, Faculty of Arts, Lagos State University.
`olawunmioni1@yahoo.com / Olawunmi.oni@lasu.edu.ng`

1 Introduction

Common ground relates to "sharedness" which should usually exist between interlocutors for them to operate on the same wavelength. Participants in any conversational encounter operate on the assumption that there are certain aspects in their encyclopedic knowledge that should be a common denominator in assessing understanding and interpreting any ongoing subject. Common ground is 'co-constituted' by the participants in any given instance of language use (cf Stalnaker, 2002; Enfield, 2008; Jaszczolt, 2008). The importance of common ground in tertiary institution students' sexual discourse (henceforth TISSD) will suffice in this paper.

2 Methodology

Forty purposive tape recordings of students' conversations were made in two universities. Two hundred structured interviews were conducted with 50 students in each of the institutions, and four focus-group discussions were held with six students each in the institutions. Participant observation was randomly undertaken on the students' interactions. The corpus was examined for the linguistic and pragmatic resources inherent in the students' interactions using a mix of pragmatic act theory, a contextual belief model and cognitive metaphor theoretic elements.

3 Theoretical Orientation

This paper benefits from Mey's (2001) Pragmeme, Odebunmi (2006) Contextual belief model and Conceptual metaphor theory.

3.1 Pragmatic Act theory

Mey's (2001) theory of pragmeme consist two broad categories: activity part and textual part. The activity type encompasses possible acts: speech acts, indirect speech acts, conversational acts etc that can be performed by interactants. The textual part covers both textual and contextual considerations. These are captured in the features: INF (inferencing), REF (reference), REL (relevance), VCE (voice), SSK (shared situational knowledge) and MPH (metaphor). M refers to any "metapragmatic" element that surfaces on the text and that directs our attention to something beyond the text – something on the "metapragmatic plane" (Mey 2001:221).

3.2 Contextual belief model

Odebunmi (2006) explains that "beliefs or assumptions held prior to or during occasions of interaction come into and facilitate the communicative process." Basically, there are two levels of beliefs: language level and situation level. Shared contextual beliefs as highlighted by Odebunmi include shared knowledge of topic, shared knowledge of word choices, referents, and references, and shared socio-cultural experiences, previous or immediate. We shall adopt Odebunmi's idea of shared cultural knowledge (SCK) in addition to Mey's shared situation knowledge (SSK) in this paper.

3.3 Conceptual metaphor theory

Conceptual metaphor theory is one of the contemporary metaphor theories; others include mental space theory, frame semantics, cognitive blending theory, metaphor power theory, space discourse theory and Lexical concept of cognitive metaphor. In cognitive mapping, correspondences are made between domains in terms of structural, ontological and orientation mappings. Kovecses (2006) explains that variation in the use of metaphors can occur along a number of dimensions including social, regional, ethnic, style, sub-cultural, diachronic and individual dimensions. In essence, understanding metaphoric language use, especially as it relates to TISSD entails taking all of these dimensions into perspective.

4 Findings

Using our theoretical orientation, we have grouped the observed common ground in TISSD into three: shared cultural knowledge (SCK), shared situational knowledge (SSK) and shared experiential knowledge (SEK).

4.1 Shared cultural knowledge (Shared Knowledge of culture-tainted slang)

Slangs are used in TISSD on the basis of shared belief that explicit mention of taboo concepts is regarded as immoral in mainstream Nigerian culture.

- Shared knowledge of Indigenous language expressions" e.g apako and kerewa
- Blending of foreign and indigenous language words**:** Chickala, chickito
- Foreign language words used differently: collabo
- Shared knowledge of indexicals. Examples include: there's this girl (in my hostel), (there's) this guy, the guy, that uncle, that my baby, that chikala etc. By so doing, outsiders lack the initial reference and also lack the grounding for inferring.

4.2 Shared situational knowledge

This is indexed by the students' use of slang words, indexicals and ellipsis. Here, we discuss only the use of indexicals:

4.2.1 Indexicals

Unclear antecedents (it), underspecified descriptions (their thing).

4.3 Shared Experiential knowledge

This bifurcates into Shared personal co-experiential knowledge and Shared extrapersonal co-experiential knowledge. We discuss them in turn:

4.3.1 Shared personal co-experiential knowledge

This feature in the following ways: use of attitudinal markers, anticipatory completion and anticipated utterance clue because as far as the narration is concerned they can also be 'potential tellers' (Liddicoat 2011). We discuss use of attitudinal markers only:

Attitudinal markers

The attitudinal markers that are identified in TISSD include "you know", "yeah/yes. These are also discourse markers but they function pragmatically as attitudinal markers.

4.3.2 Shared extrapersonal co-experiential knowledge

This features in the use of linguistic and cognitive mapping. Linguistic mapping involves using words arbitrarily to match any aspect of sexual discourse just because the words have some sort of resemblance or sound with the source domain. Cognitive mapping, on the other hand, only uses the experience from one domain to match another in order for the experience to be vividly understood by the hearer. In TISSD, the vagina is conceptualised as a house. It has the feature of a door.

5 Conclusion

In TISSD, common ground trifurcates into SCK, SSK and SEK which are characterised by the students' coded use of metaphors, slang words, indexicals, linguistic and cognitive mappings. Thus, understanding the language of sex among the students requires background knowledge of the social, linguistic and culture-specific interactional resources the students draw upon in their sexual discourse.

Appendices- Sample of Conversations

Extract 1

A: I'm not lying! I'll tell you, I like playing, Ere ipa (Rough play), Ma fun e lese(I'll give you blow..
B: So you were doing that too?
A: Seriously, so from there before you know it, eye contact and all, then „gen gen!"
C: Gen gen! gen gen!
@
A : gen gen /gen gen/, that very thing!

Extract 2

A: Umugi(/humugi/) is when uncle head me badly!
B: Yeah, when he nods you.

Extract 3

A: when we finally got home, we talked, we were talking
B: On that particular day!
A: On that particular day, we were talking about experiences, you know, me I like playing o

References

Enfield, N.J. 2008. Commonground as Resource for Social Affiliation. *Intention, Common ground and the Egocentric Speaker-Hearer.* New York,Mouton de Gruyter. 223-254.

Kovecses, Z. 2006. *Language, Mind and Culture: A Practical Introduction.* Oxford, Oxford University press.

Mey, J.L. 2001. *Pragmatics.An Introduction* Oxford,Blackwell Publishing.

Odebunmi, A. 2006. *Meaning in English: An Introduction*. Ogbomoso, Critical Sphere.

Palmer, F.R. 1985. *Semantics*. Cambridge, Cambridge University Press.

Jaszczolt, K.M. 2008. Psychological Explanations in Gricean Pragmatics and Frege's Legacy.*Intention, Common ground and the Egocentric Speaker – Hearer.* I.Kecskes and J. Mey. Eds. New York, Mouton de Gruyter. 9-44.

Stalnaker, R. 2002. Common ground. *Linguistics and Philosophy* 25: 701-721.

Ludicoat, A.J. 2011. *An Introduction to Conversation Analysis.* New York, Continuum.

Toward a semantics for French short positive feedback utterances

Laurent Prévot
Aix Marseille Université & CNRS, Laboratoire Parole et Langage
5 Av. Pasteur, 13100 Aix-en-Provence, FRANCE
laurent.prevot@lpl-aix.fr

Although feedback utterances are ubiquitous in dialogue and identified as a crucial aspect of language interaction, existing semantic/pragmatic proposals do not offer a comprehensive model for them. The present abstract argues for a two-step formal model using fairly standard semantics for lexical item entries and attempting to derive the pragmatic communicative functions from the semantics thanks to rich context modeling. The objective is not only the formal model *per se* but the possibility of using it as the backbone for a more empirical approach, in the spirit of (Gravano et al., 2012) or (Neiberg et al., 2013) but in which the usefulness of the semantic layer will be investigated.

Phenomena targeted Following the original proposal of (Yngve, 1970), we take *back-channel utterances* as utterances produced on the back channel of the conversation. If they were produced on the main channel, they will disrupt the flow of the speaker at that moment. Following (Bunt, 1994), we take *feedback utterance* as an utterance through which a dialogue participant provides information about his processing of the partner's previous utterances. This includes information about perception, interpretation, evaluation (agreement, surprise,...) or dispatch (fulfillment of a request,...). The topic of this paper are the *positive* feedback.[1]

Objectives and related work A crucial objective for our formal model is to help make more precise the interaction between the different modalities involved. Our starting point is a model in which all the feedback utterance instances associate a base form,[2] a prosodic form[3] and more acoustic-phonetic properties. When visual channel is involved, gestures or facial expressions can be combined and/or constitute another type of base forms. In previous empirical work (Stolcke et al., 2000; Gravano et al., 2012; Neiberg et al., 2013) all the instances received a communicative function. However, little has been said precisely in term of semantics. In the formal work of (Bunt, 2012), there is room for semantics but the proposal kept this part relatively abstract. Consequently, this work also aims at reducing the current gap between theses data-drivenstudies and more theoretical contributions such as (Ginzburg, 2012) or (Lascarides and Asher, 2009).

(Stolcke et al., 2000; Gravano et al., 2012; Neiberg et al., 2013) are all proposing some conclusions about the impact of the different features in their learning or classifying systems but the results are semantically shallow. In (Stolcke et al., 2000) and (Gravano et al., 2012), we only learn (on this aspect) that the stronger clues are the tokens identity which is the shallow way to get to semantics. (Gravano et al., 2012) shows the relevance of positional features which are shallow discourse features but do not enter in their analyses. (Neiberg et al., 2013) has the more surprising results that the base form is not relevant and it is phonological operations and prosody that are contributing the more to the communicative function.[4]

Providing a semantics for lexical forms, for phonological prosodic forms (contours) and for facial expressions will allow to study precisely how their interact and whether they behave more like in a compositional or a constructional fashion. This question will be addressed within and across modalities. More precisely, it is hypothesized that having a formalization of the pragmatic impact (in a given context) of a given lexical/prosodic/gesture association and a formal semantics for each of these elements we will be able to understand how they combine. A preliminary and easier to answer question in which semantics can help also is whether all these forms are compatibles. Predicting incompatibilities from our

---

[1] Of course, the polarity of the item considered is not a good clue for the positive discourse function since, for example, *'no'* is regularly used as positive feedback targeting negative utterances.

[2] Here we restrict the study to a closed list of lexical items and their combinations or repetitions.

[3] One issue is however that phonological categories prove to be very difficult to annotate on these rather reduced forms (D'Imperio et al., 2013).

[4] However, this could be due to functions there are looking at and to the way the specifically select the data subset for their experimental study.

model will be an interesting intermediate check for our approach.

Selection of forms The selection of the forms studied is straightforward. There are the most frequents forms found in our French spoken corpora. The seven tokens *ouais, oui, d'accord, voilà, okay, mh, ah* represents almost 10% of the total number of tokens in a MapTask Corpus (Gorish et al., 2014) and nearly 6% in a French conversational corpus (Prévot et al., 2013). Other potential feedback items are very far in term of frequency from the set we plan to scrutinize in this work.

The proposal Space constraints prevent a detailed presentation of the model, but overall we treat the lexical items in a relatively standard way either as propositional adverbs or type $\langle t, t\rangle$ for *oui, ouais* or as attitudinal operators of type $\langle\langle e, t\rangle, t\rangle$ for *d'accord, okay, voilà, ah*. *mh* is a special case since we consider that in default use it does not target propositional content and we will detail further its case here. Prosodic contours are also expected to act as operators on the propositional content that hold them.[5] Finally, gestures can be both modeled as propositional or attitudinal operators.

A key issue is to handle potential redundancy across modalities. The solution proposed at this stage is to combine the different attitudinal contributions through an unification-based mechanism. The issue become therefore to identify the relevant sets of dimension.

Illustration The first step of the work is to propose semantic entries for each of the forms considered. Lacking space for a definition of each attitude we try to provide explicit labels for a subset of attitudes.

Lexical items *ouais,oui*: $\lambda P.P$
ah: $\lambda P.x\ attitudeSet(x, P) \sqcup surprised$
ok,d'accord: $\lambda P.x\ attitudeSet(x, P) \sqcup agree$[6]
voilà: $\lambda P.x\ attitudeSet(x, P) \sqcup manifest$[7]

Prosodic contours Both contours types ($\sim$: high F0 standard deviation ; $\searrow$: Fall ; $\nearrow$: Rise) and their meaning are proposed, based on previous literature of several languages, for illustrating the model and might not reflect what they will be ultimately in the French case.
$\sim$: $\lambda P.x\ attitudeSet(x, P) \sqcup surprised$
$\searrow$: $\lambda P.x\ attitudeSet(x, P) \sqcup grounded$
$\nearrow$: $\lambda P.x\ attitudeSet(x, P) \sqcup elicit$

Gestures Same comment as for prosody.
☺-NOD: $\lambda P.x\ attitudeSet(x, P) \sqcup grounded$
☺-SMILE: $\lambda P.x\ attitudeSet(x, P) \sqcup amused$

The research objective can be therefore formulated as characterizing the $\sqcup$ operations (including when attitudes are conflicting). As mentioned above, a first step consists in checking the incompatibilities. Here for example, *agree* is not compatible with *elicit* and therefore *d'accord* should not be compatible with $\nearrow$.

References

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

Bunt, H. (2012). The semantics of feedback. In *16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012)*, pages 118–127, Paris (France).

D'Imperio, M., Petrone, C., and Prévot, L. (2013). Basic prosodic transcription of short french feedback utterances. In *Proceedings of Second Advancing Prosodic Transcription for Spoken Language Science and Technology*, Lisbon.

Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Gorish, J., Astésano, C., Bard, E., Bigi, B., and Prévot, L. (2014). Aix map task corpus, the french multimodal corpus of task-oriented dialogue. In *submitted*.

Gravano, A., Hirschberg, J., and Beňuš, Š. (2012). Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1):1–39.

Lascarides, A. and Asher, N. (2009). Grounding and correcting commitments in dialogue. *Journal of Semantics*.

Neiberg, D., Salvi, G., and Gustafson, J. (2013). Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*.

Prévot, L., Bigi, B., and Bertrand, R. (2013). A quantitative view of feedback lexical markers in conversational french. In *Proceedings of SIGDIAL*.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the chicago linguistic society*, pages 567–578.

---

[5] This is not a general proposal for the meaning of prosody.

[6] Actual grounding can be derived through the properties of the attitudes (eg. $grounded(x, P) \rightsquigarrow P \in gameBoard_x$)

[7] P is manifest for x

The meaning of French H*L%-contour

Uwe Reyle
IMS, University of Stuttgart
LPL, Aix en Provence
uwe.reyle@ims.uni-stuttgart.de

Cristel Portes
Laboratoire Parole et Langage,
Aix en Provence
cristel.portes@lpl-aix.fr

Abstract

The project presented here aims at developing a fine grained semantic-pragmatic analysis of the contribution of intonational contours to utterance meaning within the framework of Discourse Representation Theory (DRT). We focus on the French contour H*L% since it conveys the idea of a potential disagreement between the interlocutors.

Assertive and interrogative mode of utterances are complementary in that the speaker aims at conveying information when uttering an assertion and asks for information by means of a question. Recent work on inquisitive semantics has shown that assertions may also bear inquisitive components (Groenendijk and Roelofsen, 2009), and rising intonation in declarative questions expresses the speakers commitment to a proposition but at the same time mark it as contingent on ratification by the addressee (Gunlogson, 2008).

French intonation is particularly rich in contours that provide even more illocutionary facets for interpretation than English (Beyssade and Marandin, 2006; Portes and Reyle, 2013; Portes and Beyssade, to appear). Portes and Reyle (2013) follow Krifka's (to appear) proposal to interpret speech acts by development of spaces of commitments assigned to the discourse participants. The meaning components Portes and Reyle (2013) attribute to the four French contours they are analyzing are (i) preconditions on the hearers previous commitments, (ii) speaker's own commitments, and (iii) expectations concerning the hearer's commitments as a result of the processed utterance. Uttering ϕ with a rising-falling contour (H*L%), for example, has an assertive component (the speaker commits himself to ϕ and asks the hearer to accept ϕ as well), but it also presupposes that the hearer is publicly committed to the negation of the utterance, and that in addition the speaker believes that the hearer should know that ϕ holds.

The present contribution aims at a representational format of utterances in DRT that does justice to these subtle distinctions along the following lines.

A pure assertion of p = *Lola smokes.* (realized with a falling contour) will be represented by a DRS like (1).[1]

$$(1)\qquad \begin{array}{|c|}\hline \text{s e} \\ \hline \text{e:utter(sp,p)} \quad \text{e} \subseteq \text{s} \\ \text{ISSUE}\{\text{p},\neg\text{p}\} \\ \text{s:(sp,}\left\{\begin{array}{l} <\text{EVI,p}> \\ <\text{DES,s}'':\text{know(h,p)}> \end{array}\right\}) \\ \hline \end{array}$$

The first two conditions say that the speaker, sp, has uttered p during a state s. This state is an attitudinal state of the speaker which is described in the last condition. It consists of two parts: (i) that the speaker bears some evidential relation to p, i.e. p is positively anchored by some causal relation to the proposition it represents (<EVI,p>), and (ii) that he has the desire that the hearer, h, knows whether p is true or not (<DES,s'':know(h,{p})>). The condition <EVI,p> is underspecified wrt. what particular type of evidential relation is involved. We have <EVI,p> ⇔ (¬WON,{p,¬p})> ∧ ¬CONFL,{p,¬p})>), i.e. S is neither in a state of conscious ignorance with respect to p nor is he in a state of internal conflict wrt. it.[2] Furthermore different dimensions of evidential relations are to be distinguished: (i) in which part of the speaker's articulated context (viz. (Kamp,)) p is anchored, and (ii) which source and type (witness, hear, surmise, (viz. (Murray, 2013))) of evidence it has. The second condition on (1) presents the issue, i.e. a set of propositions awaiting resolution through

---

[1] We ignore details about tense.

[2] This means that he does not have both, an anchor for p and another anchor for ¬p.

selection of one or more of its members.[3] A sentence settles an issue if it contains a focussed constituent that gives rise to an alternative set which corresponds to this issue. After the assertion of p the hearer is under the obligation to settle the raised issue, ISSUE{p,¬p} by either accepting or rejecting p.

With a rising contour an utterance of *Lola smokes?* is a declarative question. This question is similar to the assertion in that it rises the issue whether p or not, and in that it may be assumed that the speaker has evidence for p. The difference to the assertive utterance is that, now, the desire of the speaker is to know himself whether p or ¬p holds, i.e. $<DES,s'':know(sp,\{p,\neg p\})>$ instead of $<DES,s'':know(h,p)>$. Note that the speaker's evidence may depend on the hearer (e.g., when he echoes a previously made assertion in confirmation questions) or that it may be rooted hearer-independly. Polar questions (with questioning syntax) lack a reference to the speaker's evidence for p.

Yes/no-answers and *aha*-responses differ wrt. the speaker's evidence for p/¬p. In the case he answers with *yes* he must himself have some evidence for p and with *aha* he signals that he will add p to his positive beliefs, together with an internal anchor for the source of the information. The negative answer, *no*, is the speaker's assertion of ¬p.[4] It presupposes a prior move, an assertion or a question, and answers its ISSUE. In both cases the negative answer must be based on the speaker's evidence for ¬p. As response to previous assertions a conflict in the commitment space results: the speaker of the answer (sp) wants the hearer (h) to know ¬p and h, who has risen the presupposed issue, wanted sp to know p. (2.a) shows the presupposition and (2.b) the assertive part of a negative answer.

(2) a. $$\begin{array}{|c|}\hline p \\ \hline \text{ISSUE}\{p,\neg p\} \\ \hline \end{array}$$

b. $$\begin{array}{|c|}\hline e \\ \hline e:utter(sp,\neg p) \\ s:(sp,\left\{\begin{array}{l} <EVI,\neg p> \\ <DES,s':know(h,\neg p)> \end{array}\right\}) \\ \hline \end{array}$$

French declarative utterances with an H*L%-contour are like full negative responses to ISSUEs created by assertions, i.e. they presuppose the issue whether p or ¬p. But in addition they presuppose that the hearer, h, has a propositional attitude which corresponds to his assertion of ¬p. These presuppositions are shown in (3.a). Hence they create the same conflict as *no*-answers. But in addition they indicate a way out of the conflict by addressing the evidential state of h conveying in addition: *you should know*, represented by $s_3:(h,\{<EVI,p>\})$ in the assertive part (3.b) of the representation.

(3) a. $$\begin{array}{|c|}\hline s_h\ p \\ \hline \text{ISSUE}\{p,\neg p\} \\ s_h:(h,\left\{\begin{array}{l} <EVI,\neg p> \\ <DES,s_1:know(sp,\neg p)> \end{array}\right\}) \\ \hline \end{array}$$

b. $$\begin{array}{|c|}\hline e\ s_{sp} \\ \hline e:utter(sp,p) \\ s_{sp}:(sp,\left\{\begin{array}{l} <EVI,p> \\ <DES,s_2:know(h,p)> \\ <BEL,s_3:(h,\{<EVI,p>\}> \end{array}\right\}) \\ \hline \end{array}$$

In (3) the presupposed issue is settled by the assertive part of the representation. In addition the attitudinal state of the hearer wrt. ¬p and the corresponding parts of the attitudinal state of the speaker wrt. p form a set of alternatives that allows us to consider the contribution of the H*L%-contour as marking a contrast in the sense of there being a clearly defined set of alternatives to the asserted part. This view is supported by lots of examples we have analyzed in the SID-corpus (Bertrand et al., 2008), as e.g., in *C'est des chataignes*$^{H*L\%}$ *bien sûr*$^{H*L\%}$ *ouais il y a que ça*$^{H*L\%}$ *qui est comestible.*, where the first contour contrasts with anybody's (except the speaker's) potential assumption that it's not *des chataignes*, but *des marrons*, the second contrasts the different types of evidential relations and the third reinforces the contrast already expressed by the first. It is important to note that contrast marking co-occurs here with cleft and pseudo-cleft constructions that are necessary in French to mark focus.

---

[3] Issues may be raised by assertions or by asking explicit questions, but in many situations they arise implicitly. Following (Roberts, 1996), (Büring, 2003), (Guinzburg, 2012) we may assume that the set of outstanding issues at any given stage in a discourse is arranged in a stack-like structure from which issues are removed when they are settled and to which new issues may be added.

[4] Note that assertions do not introduce a discourse referent for ¬p. We analyze *yes/no*-answers by structured propositions and in analogy to short answers to constituent questions. To this end polar issues should be represented in the form ISSUE(< $\begin{array}{|c|}\hline Q \\ \hline Q \in \{\lambda K\,.\,K, \lambda K\,.\,\neg K\} \\ \hline \end{array}$, Q(p)>).

References

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 (TAL)*, pages 105–134.

Claire Beyssade and Jean-Marie Marandin. 2006. French intonation and attitude attribution. In P. Denis, E. McCready, A. Palmer, and B. Reese, editors, *Proceedings of the 2004 Texas Linguistics Society Conference: Issues at the Semantics-Pragmatics Interface*.

Daniel Büring. 2003. On d-trees, beans, and b-accents. *Linguistics & Philosophy*, 26:5:511–545.

Jeroen Groenendijk and Floris Roelofsen. 2009. Inquisitive semantics and pragmat- ics. In *Workshop on Language, Communication, and Rational Agency*. Stanford.

Jonathan Guinzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Christine Gunlogson. 2008. A question of commitment. In De Brabanter, Philippe, and Patrick Dendale, editors, *Commitment*, pages 101–136.

Hans Kamp. Discourse structure and the structure of context. IMS, Univesity of Stuttgart.

Manfrd Krifka. to appear. Negated polarity questions as denegations of assertions. In F. Kiefer and C. Lee, editors, *Contrastiveness and scalar implications*. Springer.

Sarah Murray. 2013. Varieties of update. *Semantics and Pragmatics*.

Cristel Portes and Claire Beyssade. to appear. Is intonational meaning compositional? *Verbum*.

Cristel Portes and Uwe Reyle. 2013. Intonational meaning triggers expectations. In *Discourse Expectations: Theoretical, Experimantal and Computational Perspectives (Detec2013)*, Tübingen.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Ohio State University Working Papers in Linguistics*, 49.

The contribution of speech-rhythm to end-of-utterance detection

Franziska Schaller
Experimental Neurolinguistics Group
Faculty of Linguistics and Literary Studies
Bielefeld University
P.O. Box 100131
33501 Bielefeld, Germany
`franziska.schaller@uni-bielefeld.de`

Horst M. Müller
Experimental Neurolinguistics Group
Faculty of Linguistics and Literary Studies
Bielefeld University
P.O. Box 100131
33501 Bielefeld, Germany
`horst.mueller@uni-bielefeld.de`

Abstract

In the current study we focused on the question of how language specific properties other than syntax and semantics – especially rhythm – affect end-of-utterance detection. We compared the anticipation timing accuracy for German stimuli (mother-tongue) to the timing accuracy for foreign language items and sinusoidal tones. Subjects were more accurate when anticipating the ends of utterances in stress-timed than in syllable-timed languages or tones. We interpret these results as an indicator for rhythm being relevant in end-of-utterance anticipation.

1 Introduction

Interlocutors show accurate timing in conversation. This has already been suggested by Sacks et al. (1974), who developed an established turn-taking system. From a projectionists' point of view, recipients anticipate when a speaker's turn ends, which permits very precise speaker changes. For this anticipation process, lexico-syntactic characteristics seem to be particularly relevant (e.g. Beattie, 1981; de Ruiter et al., 2006; Magyari et al., 2011; Müller et al., 2013). Other studies suggest that prosodic and suprasegmental characteristics are important (e.g. Gravano and Hirschberg, 2009; Heldner et al., 2009; Wells and Macfarlane, 1998). Also, it is discussed whether successful turn-taking processes depend on the interplay of several aspects – including e.g. semantics, syntax, prosody, and rhythm (e.g. Ford and Thompson, 1996; Selting, 1996).

As for rhythm, Wagner et al. (in press) suggest that there is a strong connection between prosodic characteristics of an utterance and speech rhythm, insofar as the prosodic features follow a regular oscillation pattern. Interlocutors entrain their speech rhythm according to this pattern (Inden et al., in press) and could use it to adjust their turn onsets (Auer et al., 1999; Couper-Kuhlen, 1993; Wilson and Wilson, 2005). However, there is no empirical evidence for this hypothesis so far (Inden et al., in press).

The intention of the current study was to assess whether speech rhythm and general articulatory speech-specific features other than syntax and semantics are relevant for end-of-utterance anticipation. For that purpose, we surveyed how well participants were able to anticipate the ends of utterances in different languages and measured their anticipation timing accuracy (ATA) as an indicator of conscious behavioral processes.

2 Material and method

2.1 Stimuli

We used spoken sentences (161 total) as stimuli. There were 23 items in each of seven languages (German, English, Italian, Polish, Turkish, Arabic, Korean). Languages other than German (L1) and English (L2) were judged as unknown. Sinusoidal tones (10 total) were used as control items. The tones were generated at 450 Hz and matched the length of the sentences.

As for the unknown language utterances, we expected that participants would have to use speech rhythm and other suprasegmental features for a successful anticipation since they could not rely on semantic and syntactic content. If they do not use other elements besides syntax and semantics in their daily turn-taking, their ATA should not be much better for the linguistic, but incomprehensible stimuli as it is for the maximally non-linguistic sinusoidal tones that do not contain any linguistic information at all.

2.2 Procedure

The items were presented auditorily (45 to 55 dB) and subjects listened to them with headphones. They were asked to push a button on an external response box at the exact moment the utterance ended. The time span between the actual end of the utterance and the button push was defined as the ATA.

3 Results

In addition to checking for ATA differences between the languages, we also grouped them as a) either Indo-European (IE) or Non-Indo-European (Non-IE) and b) either stress-timed or (rather) syllable-timed. Both the IE and the stress-timed group contained the stimuli in the known languages.

Comparisons of the ATA of the foreign language stimuli and the tones revealed several differences. As expected, subjects reached a better ATA for the ends of German items than for any other stimulus type. Further, they anticipated the ends of tones and of Turkish stimuli equally worse than the ends of all other stimulus types. A repeated measures ANOVA ($F(3.42, 119.52) = 100.27$, $p \leq .001$) and the Bonferroni multiple comparison post-hoc test showed that there were significant differences between the ATA of almost all item types. The comparison of Polish, Turkish and Korean items to tones revealed no significant differences in the ATA.

As well, there was an overall highly significant effect ($F(1.33, 46.65) = 98.35$, $p \leq .001$) when comparing the ATA of IE languages to that of Non-IE languages and of tones. All stimulus groups differed significantly from each other. The ends of IE utterances were most accurately anticipated. Since these results suggest that there must be some suprasegmental elements relevant for end-of-utterance anticipation, we grouped the sentence types according to their stress pattern. The ATA differences between stress-timed and syllable-timed languages and tones were highly significant ($F(1.35, 47.32) = 116.61$, $p \leq .001$). The ends of stress-timed utterances were anticipated significantly better than of syllable-timed items ($p \leq .01$) and of tones ($p \leq .01$).

4 Discussion

The ATA differences between foreign language stimuli and tones were mostly not significant, which implies that anticipation performance was definitely better when subjects had access to semantics and syntax and that suprasegmentals alone were not sufficient for an adequate anticipation performance. Nevertheless, subjects anticipated the ends of the Non-IE utterances better than of tones although they did not have access to syntax and semantics. It is probable that they used language-universal linguistic properties – which we suppose to be suprasegmental in nature – to anticipate the ends of utterances in unknown languages. Possible relevant properties in this context are e.g. the last major accent and specific F_0-contours that have been discussed in a number of corpus studies (Caspers, 2003; Heldner et al., 2009; Koiso et al., 1998; Wells & Macfarlane, 1998). Further, language differs from tones in its speech specific rhythm, which might be relevant in the anticipation of utterance-ends as well (e.g. Beňuš et al., 2011). Our results support this assumption. There was a significant difference between the ATA of stress-timed vs. syllable-timed languages. Thus, participants were more accurate when anticipating the ends of Arabic items, the stress pattern of which is similar to that of German, than when anticipating the ends of e.g. Polish items that differ from German in their stress pattern. Further, there were no significant differences between the ATA of utterances with a syllable-timed rhythm and the ATA of tones. Thus, subjects' anticipation performance was inadequate when they were not able to make use of neither syntax and semantics nor a well-known stress pattern. This implies that rhythm probably is an important feature that people use when anticipating the end of an utterance.

5 Conclusion

We propose that in combination with syntax and semantics, rhythm is a relevant characteristic in the anticipation of utterance ends. As well, there are other suprasegmental characteristics which influence anticipation processes, albeit to a lesser degree than syntax and semantics.

The results of the current study thus support the view that there is a number of features that are all accounted for by conversational partners when anticipating the end of a turn.

Acknowledgments
This study was supported by the DFG (SFB 673). We thank Jan de Ruiter and Petra Wagner for their helpful comments.

References

Augustín Gravano and Julia Hirschberg. 2009. Turn-yielding cues in task-oriented dialogue, in *Proceeding of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 253-261, eds. Patrick G. T. Healey, Roberto Pieraccini, Donna K. Byron, Steve Young, and Matthew Purver. Association for Computational Linguistics, Stroudsburg, PA.

Benjamin Inden, Zofia Malisz, Petra Wagner, and Ipke Wachsmuth. In press. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent, in *Proceedings of the 15th International Conference on Multimodal Interaction, ICMI '13*, eds. Julien Epps, Fang Chen, Sharon Oviatt, Kenji Mase, Andrew Sears, Kristina Jokinen, and Björn Schuller. ACM, New York, NY.

Bill Wells and Sarah Macfarlane. 1998. Prosody as an interactional resource. Turn-projection and overlap. *Language and Speech*, 41: 265-294.

Celia E. Ford and Sandra A. Thompson. 1996. Interactional units in conversation. Syntactic, intonational and pragmatic resources for the management of turns, in *Interaction and grammar*, 134-184, eds. Elinor Ochs, Emanuel A. Schegloff, and Sandra A. Thompson. Cambridge University Press, Cambridge, UK.

Elizabeth Couper-Kuhlen. 1993. *English speech rhythm: Form and function in everyday verbal interaction.* Benjamins, Amsterdam.

Geoffrey W. Beattie. 1981. The regulation of speaker turns in face-to-face conversation: Some implications for conversation in sound-only communication channels. *Semiotica*, 34: 55-70.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, 41: 295-321.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50: 696-735.

Horst M. Müller, Stefanie Jansen, and Hendrik Wesselmeier. 2013. Anticipation of speakers' end-of-turn indicated by lateralized readiness potential. *Society for Neuroscience*, 11403.

Janneke Caspers. 2003. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, 31: 251-276.

Jan P. de Ruiter, Holger Mitterer, and Nick J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82: 515-535.

Lilla Magyari, Marcel C. M. Bastiaansen, Jan P. de Ruiter, and Stephen Levinson. 2011. *Neuronal correlates of anticipation related to turn-taking in conversations.* Poster presented at the AMLaP 2011, Sept 1st-3rd, Paris, France.

Margret Selting. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, 6: 357-388.

Margaret Wilson and Thomas P. Wilson. 2005. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, 12: 957-968.

Mattias Heldner, Jens Edlund, Kornel Laskowski, and Antoine Pelcé. 2009. Prosodic features in the vicinity of silences and overlaps, in *Nordic Prosody: Proceedings of the Xth Conference, Helsinki 2008*, 95-105, eds. Martti Vainio, Reijo Aulanko, and Olli Aaltonen. Lang, Frankfurt.

Peter Auer, Elizabeth Couper-Kuhlen, and Frank Müller. 1996. *Language in time*. Oxford University Press, Oxford.

Petra Wagner, Zofia Malisz, Benjamin Inden, and Ipke Wachsmuth. In press. Interaction phonology – A temporal coordination component enabling representational alignment within a model of communication, in *Advances in interaction studies. Alignment in communication*, 109-132, eds. Ipke Wachsmuth, Jan P. de Ruiter, Petra Jaecks, and Stefan Kopp. Benjamins, Amsterdam.

Štefan Beňuš, Augustín Gravano, and Julia Hirschberg. 2011. Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, 43: 3001-3027.

Successful Strategies for Ambiguity Resolution in Dialogue

Gesa Schole
RTG 1808 Ambiguity
University of Tübingen
gesa.schole@uni-tuebingen.de

Thora Tenbrink
School of Linguistics and English Language
Bangor University
t.tenbrink@bangor.ac.uk

Elena Andonova
Research Centre for Cognitive Science
New Bulgarian University
eandonova@nbu.bg

Kenny Coventry
School of Psychology
University of East Anglia
k.coventry@uea.ac.uk

1 Introduction

Everyday communication is characterised by the common phenomenon of ambiguity (Winter-Froemel and Zirker, 2010), which occurs when more than one meaning is associated with one item (Ziegler, 2010). Many spatial terms, for example, can be interpreted in different ways and are thus inherently ambiguous (Schober, 1993).

In dialogue, speaker and addressee must agree on one of the potential interpretations to enable understanding. Our study addresses referring strategies that interlocutors use to specify object location, and associated problems that may cause a failure of object placement. We present a qualitative analysis of the negotiation of a bedside table's location as a case study for object placement in dialogue, contrasting the description of functional and non-functional spatial arrays.

2 Empirical Study

Our dialogue corpus (first reported in Tenbrink et al., 2008) was collected using a spatial reference task between two participants who were uninformed about the research goals. One of them (the *director*) had a furnished four-room doll's house in front of them; their task was to instruct the other one (the *matcher*) to furnish another (empty) doll's house in the same way. Verbal communication was not restricted, but the participants could not see each other. Conditions differed as to the arrangement of furniture (fig. 1). In the **functional** condition, rooms represented the functions of bathroom, bedroom, living room, and kitchen. In the **non-functional** condition, the furniture pieces were arranged randomly. Speakers often use functional features in spatial descriptions (Andonova et al., 2010).

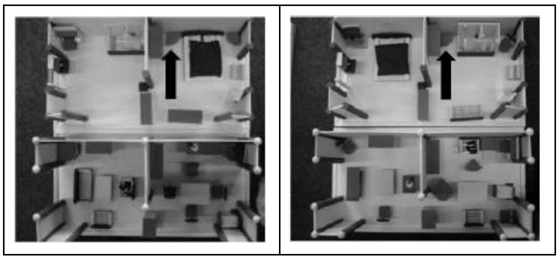


Figure 1: The functional (left) and the non-functional condition (right). Arrows mark the bedside table.

3 Results

Out of the corpus, in the following we examine how eight randomly selected dyads in each condition negotiated the location of a bedside table that had the same position in both arrays.

In the **functional** condition, in six of eight cases the bedside table was positioned correctly. One placement error occurred as a consequence of the failed negotiation of the previous object, but the negotiation of the table itself was consistent and unproblematic. In these seven successful cases, the table was introduced in terms of a cluster (functional group) together with another bedside table and a bed between the two tables. The following exemplifies this:

(1) *director*: äh rechts und links vom Ehebett auch an der Wand stehen so Nachttischschränke, [uh to the right and left of the bed and against the wall there are sort of bedside tables] (…)
matcher: die stell ich? [I put them?]
director: die stellst du links und rechts vom Bett auf [you put them to the left and right of the bed]

Clustering objects implies that the furniture pieces share functionality, with the effect that the relations between the clustered objects may be inferred from world knowledge.

The only failed negotiation in the functional condition that was not a follow-up error did not use clustering:

(2) *matcher*: und der steht dann jetzt direkt an dem Schrank dran? [and it is now placed directly against the wardrobe?]
director: genau, so daneben dann. [exactly, sort of beside it.]

This information about the object's location is ambiguous. The preposition *daneben* (beside) requires a specification of perspective, as the bedside table may be standing beside the wardrobe from the speaker's viewpoint (fig. 2, left) or from the wardrobe's viewpoint (fig. 2, right). Since perspective is implicit in example (2), the error can be traced back to a perspective discrepancy that the director and matcher did not notice.

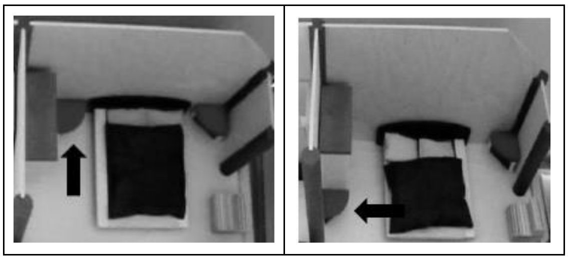


Figure 2: The bedside table beside the wardrobe: from the speaker's viewpoint (left) and from the wardrobe's viewpoint (right).

In the **non-functional** condition only four out of the eight dyads managed to place the bedside table correctly. One error depends on a similar perspective discrepancy as example (2):

(3) *director*: und das stellst du dann da so vor dass es ähm dies Runde [and then you put this in front of that so that the round part]
matcher: mhm [uhuh (affirmative feedback signal)]
director: ins Zimmer guckt das heißt ähm die beiden Ecken [faces the room that is uhm the two corners]
matcher: mhm [uhuh]
director: sind an den Wänden [are against the walls]
matcher: ja [yes]
director: einmal an dem Schrank und einmal an der Wand [one at the wardrobe and one at the wall]
matcher: ja, ja hab ich [yes, yes got it]

These instructions neither specify which wall is meant nor the perspective underlying the expression *in front of that*. The object may either be placed in front of the wardrobe from the speaker's viewpoint (fig. 2, right) or from the wardrobe's viewpoint (fig. 2, left), yielding the same ambiguity as with the term *daneben* seen above.

The remaining three location errors occur in spite of the fact that no ambiguous spatial information is given. The term *in front of* in example (4) is disambiguated by a specification of the wall against which the bedside table is placed:

(4) *director*: und vor diesem Regal steht ähm dieses, dieser kleine, wie so'n kleiner Hocker, [and in front of that shelf, there is uhm this, this small, like a small stool] (…) genau, das steht vor dem Regal [exactly, that is standing in front of the shelf] (…) das heißt die eine flache Seite is' an der Wand an der auch die Dusche steht und die andere flache Seite is' an dem Regal. [that means one of the plain sides is against the wall where also the shower is standing and the other plain side is against the shelf]

The matcher however does not take the information about the wall into account, but focuses on the first information provided by the director (*in front of that shelf*). Similarly, the remaining two matchers act on the basis of their initial assumption about the object's location and disregard the specifying information (*the back wall*).

The four successfully located tables in the non-functional condition were negotiated using complex references to neighbouring objects, the speaker's position, and the walls.

4 Discussion

In both conditions, negotiation of object placement could fail due to underspecification of underlying perspective. This kind of spatial ambiguity is a common phenomenon (Schober, 1995). To avoid miscommunication, interlocutors tend to be consistent in their perspective choice (Vorwerg, 2009), which saves the cost of discussing their choices explicitly while still being specific (Garrod and Anderson, 1987).

The remaining errors were due to the matcher disregarding relevant information. The dialogue extracts suggest that matchers had problems changing their initial assumptions about object location, even when provided precise, disambiguating information by the directors. In the examples, references to the walls as a feature of the overall environment played an important role for disambiguation. Although such reference to the environment is a frequent strategy (Carlson and Hill, 2008), our data suggest that it may be prone to being disregarded by matchers. These findings support the view that misunderstandings occur when the addressee disregards disambiguating information, or lacks contextual or world knowledge (Winter-Froemel and Zirker, 2010).

Of the successful strategies seen in our data, the clustering of objects into functional groups (where available) appeared as a frequent and very successful strategy, as it efficiently disambiguates location descriptions by implicitly relying on world knowledge.

References

Elena Andonova, Thora Tenbrink, and Kenny R. Coventry. 2010. Function and Context Affect Spatial Information Packaging at Multiple Levels. *Psychonomic Bulletin & Review* 17:575-580.

Laura A. Carlson and Patrick L. Hill. 2008. Processing the Presence, Placement and Properties of a Distractor in Spatial Language Tasks. *Memory and Cognition*, 36(2):240-255.

Simon Garrod and Anne Anderson. 1987. Saying What You Mean in Dialog: A Study in Conceptual and Semantic Co-ordination. Cognition, 27:181-218.

Michael Schober. 1993. Spatial Perspective Taking in Conversation. *Cognition*, 47:1-24.

Michael Schober. 1995. Speakers, Addressees, and Frames of Reference: Whose Effort Is Minimized in Conversations about Locations. Discourse Processes, 20:219-247.

Thora Tenbrink, Elena Andonova, and Kenny Coventry. 2008. Negotiating Spatial Relationships in Dialogue: the Role of the Addressee. *Proceedings of LONDIAL – The 12th SEMDIAL Workshop, June 2nd - June 4th, 2008*, King's College, London, UK.

Constanze Vorwerg. 2009. Consistency in Successive Spatial Utterances. In Kenny Coventry, Thora Tenbrink and John A. Bateman (eds.), Spatial Language and Dialogue: Explorations in Language and Space. Oxford: Oxford UP.

Esme Winter-Froemel and Angelika Zirker. 2010. Ambiguität in der Sprecher-Hörer-Interaktion. Linguistische und literaturwissenschaftliche Perspektiven. [Ambiguity in speaker-hearer interaction. Perspectives from Linguistics and Literary Studies.] *Zeitschrift für Linguistik und Literaturwissenschaft.* 158:76-97.

René Ziegler. 2010. Ambiguität und Ambivalenz in der Psychologie. Begriffsverständnis und Begriffsverwendung. [Ambiguity and ambivalence in psychology. Term definition and use.] *Zeitschrift für Literaturwissenschaft und Linguistik*, 158:125-171.

Demonstrating actions to a robot: How naïve users adapt to a robot's replication of goal and manner-oriented actions

Anna-Lisa Vollmer[1], **Manuel Mühlig**[2], **Katharina J. Rohlfing**[3], **Britta Wrede**[4], **Angelo Cangelosi**[1]
[1] School of Computing and Mathematics, Plymouth University, Plymouth, United Kingdom
{anna-lisa.vollmer, a.cangelosi}@plymouth.ac.uk
[2] Honda Research Institute Europe GmbH, Offenbach/Main, Germany
manuel.muehlig@honda-ri.de
[3] CITEC, [4] Applied Informatics Group, Bielefeld University, Bielefeld, Germany
kjr@uni-bielefeld.de, bwrede@techfak.uni-bielefeld.de

Abstract

Subscribing to an interactive view on robot action learning, we conducted a human-robot interaction experiment with inexperienced users. This paper describes an analysis on how the robot's replications influence participants' following demonstrations. Detailed results will be presented on the poster.

1 Introduction

Imitation learning in robotics aims at developing mechanisms for robots to learn from a human tutor's demonstrations (Schaal, 1999). In order to replicate an action correctly, the robot needs to understand a demonstrated action which implies knowing what is important about it. But how do tutors convey this information about an action? Recent research has proposed that in action learning the learner's mental states are aligned to those of the tutor in a bi-directional process (Thibault, 2011; De Jaegher et al., 2010). Subscribing to this view, tutoring a robot should really be a bi-directional endeavor in which the robot not merely observes the tutor's demonstrations, but actively takes part by giving feedback. In adult-infant tutoring interactions, the infant learner's feedback reveals information about his/her state of attention and understanding and has been found to influence the adult tutor's action demonstrations (Pitsch et al., in press). Analyzing data from a human-robot interaction study where human tutor and robot learner take turns in demonstrating and replicating actions, we investigate how naïve users modify their action demonstrations, when the robot replicates the action correctly or incorrectly.

2 Experiment

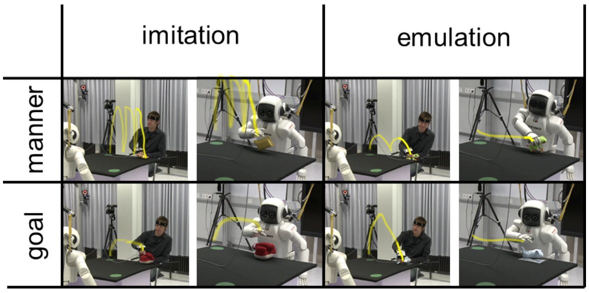


Figure 1: Experimental conditions.

The experiment was conducted with 59 participants (28 m, 31 f). Participants had no prior experience with robots and interacted with the robot for the first time.

Participants were seated at a table across from a standing humanoid robot and were asked to teach the robot eight different actions. The actions were chosen to be either goal-oriented (the end position of the involved object is the important main feature, e.g. the action 'to hang up the phone') or manner-oriented (the path is most important, e.g. the action 'to clean the window with a sponge') and the robot replicated the actions via imitation (copying the movement of the involved object as exactly as possible) or emulation (transporting the object straight to its end position) yielding correct and incorrect replication attempts. Additionally, each participant was presented with one of three robot gaze behaviors (a social gaze consisting in appropriately following the action demonstration (imitation) or anticipating the goal position (emulation condition), a random gaze consisting in various alternating gaze targets independent of the tutor's conduct and a static gaze condition in which the robot only showed a fixed gaze to the overall scene).

For each action, the participant and robot took turns in executing the action. After the participant demonstrated the action, the robot replicated it and the participant had the chance to correct the robot by demonstrating the same action again followed by another robot replication turn (for one specific action, the robot replicated the action in the same condition and did not change its behavior) and so on. After each robot repetition, the participant decided if it was necessary to demonstrate the action again or to stop and carry on to the next action.

In this analysis, we focus on the very first action each participant demonstrated to the robot because demonstrations of subsequent actions incorporated potential modifications based on experiences from previous turns. The order of actions was randomized for each participant and also if the robot imitated or emulated this action. Because goal and manner-oriented actions have very different properties, we examined the data separately for the two types of action. The analysis sets out to compare the tutor's first demonstration of the first action and the second demonstration of the same action (after the robot's replication turn). Not all participants deemed it necessary to show a second demonstration of the action leaving us with 26 participants who demonstrated a goal-oriented action (which the robot imitated in 10 and emulated in 16 cases) and 26 participants who demonstrated a manner-oriented action (which the robot imitated in 6 and emulated in 20 cases). The data used for the analysis consisted of 3D object trajectory data obtained from a magnetic-field based Polhemus Liberty System which was attached to the objects involved in the actions and tracked their movements. To compare the characteristics of the demonstrations, we computed a set of objective measures on the obtained trajectories, please refer to (Rohlfing et al., 2006; Vollmer et al.,).

3 Results and Discussion

We investigated goal-oriented and manner-oriented actions individually and conducted separate two-way mixed ANOVAs with demonstration (first, second) as within-subjects variable and robot replication condition (imitation, emulation) as between-subjects variable.

3.1 Goal-oriented actions

For goal-oriented actions, they revealed significant main effects of the demonstration for *acceleration*, and marginally significant main effects for *velocity*, *average length of motion pauses*, and *range*. Participants demonstrated the action slower in the second demonstration than in the first demonstration, and with longer motion pauses. The second demonstration was shown with less range than the first one. In summary, goal-oriented actions were shown slower and with less detail when they were repeated.

Additionally the tests revealed a marginally significant interaction effect for *action length*. T-tests as post hoc comparisons showed a marginally significant difference between the length of the first demonstration of a subsequently imitated action and the length of the first demonstration of a subsequently emulated action. One possible explanation for this finding could involve the robot gaze during the demonstrations. Indeed, when conducting follow-up tests for the three robot gaze conditions separately, we found this difference alone in the social gaze condition, which thus seems to account for the finding. Thus, the anticipating gaze of the robot in the emulation condition during the participant's action demonstration, led participants to perform the demonstration in a shorter time frame than when the robot followed the movements with its gaze in the imitation condition.

3.2 Manner-oriented actions

For manner-oriented actions, the tests revealed a marginally significant main effect of the demonstration for *width*. Participants demonstrated the action less wide at the second demonstration compared to the first one. Additionally they showed significant interaction effects for *height*, *velocity*, and *acceleration*. T-tests as post hoc comparisons showed that the second demonstration of a previously imitated action was significantly higher than the first demonstration. Results suggest that when the action was emulated by the robot, the second demonstration was of similar or even lower height than the first one. For the demonstration speed, the tests revealed that the second demonstrations of previously emulated actions were performed with less velocity and acceleration than the first ones and slower than the second demonstrations of a previously imitated action. A possible explanation for these findings is that participants exaggerate their following demonstrations to emphasize the manner of the action, when the robot successfully replicated (imitated) it. When the robot failed to replicate (emulated) the shown action, the participant showed a simpler and easier subsequent demonstration. Thus, participants adjusted their subsequent demonstrations according to the robot's capabilities and understanding witnessed in its replication as a turn-based feedback.

Acknowledgments

This research was supported by the Honda Research Institute Europe and the EU project RobotDoC (235065) from the FP7 Marie Curie Actions ITN.

References

Hanne De Jaegher, Ezequiel Di Paolo, and Shaun Gallagher. 2010. Can social interaction constitute social cognition? *Trends in cognitive sciences*, 14(10):441–447.

Karola Pitsch, Anna-Lisa Vollmer, Katharina J. Rohlfing, Jannik Fritsch, and Britta Wrede. in press. Tutoring in adult-child interaction. on the loop of the tutor's action modification and the recipient's gaze. *Interaction Studies*.

Katharina J. Rohlfing, Jannik Fritsch, Britta Wrede, and Tanja Jungmann. 2006. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199.

Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242.

Paul J. Thibault. 2011. First-order languaging dynamics and second-order language: the distributed language view. *Ecological Psychology*, 23(3):210–245.

Anna-Lisa Vollmer, Manuel Mühlig, Jochen J. Steil, Karola Pitsch, Jannik Fritsch, Katharina J. Rohlfing, and Britta Wrede. Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning. *Manuscript submitted for publication.*