Proceedings Semdial 2014, DialWatt, Edinburgh

Verena Rieser, Philippe Muller

DialWatt — Semdial 2014

The 18th Workshop on the Semantics and Pragmatics of Dialogue

Edinburgh, September 1-3, 2014

Edited by Verena Rieser and Philippe Muller

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

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

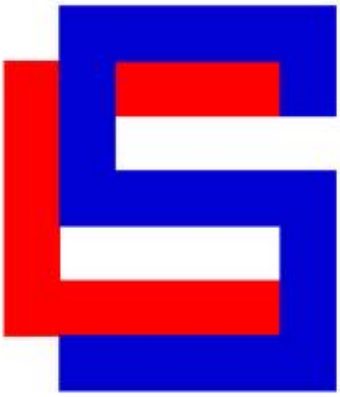
DialWatt Website

http://www.macs.hw.ac.uk/InteractionLab/Semdial/

Gold Sponsors

Silver Sponsors

Endorsements



Preface

DialWatt brings the SemDial Workshop on the Semantics and Pragmatics of Dialogue back to Edinburgh, where the sixth meeting - EDILOG - took place in 2002. The current edition of SemDial is hosted by Heriot-Watt University. The return to Edinburgh has given us opportunity to colocate SemDial with a number of other events, including RO-MAN 2014, the IEEE International Symposium on Robot and Human Interactive Communication, the RefNet summer school and workshop in Psychological and Computational Models of Language Production, and AMLaP 2014, the annual conference on Architectures and Mechansims for Language Processing.

We received a total of 31 full paper submissions, 17 of which were accepted after a peer-review process, during which each submission was reviewed by a panel of three experts. We are extremely grateful to the Programme Committee members for their very detailed and helpful reviews. In response to the call for abstracts, we received a total of 34 abstract submissions describing ongoing projects or system demonstrations, of which 32 were accepted for poster presentation.

All accepted full papers and poster abstracts are included in this volume. The DialWatt programme features four keynote presentations by Holly Branigan, Jon Oberlander, Matthew Purver and Michael Schober. We thank them for participating in SemDial and are honoured to have them at the workshop. Abstracts of their contributions are also included in this volume.

DialWatt has received generous financial support from the EU FP7 PARLANCE project, the Scottish Informatics & Computer Science Alliance, and the School of Mathematical and Computer Sciences (MACS) at Heriot-Watt University which hosts the event, we are very grateful for this sponsorship. We have also been given endorsements by the ACL Special Interest Groups: SIGdial and SIGSEM.

Last but not least we would like to thank the following people for their tireless work, Arash Eshghi who helped with all aspects of the local organisation, Mary Ellen Foster our local events organiser, and Andy Taylor for developing and maintaining the website, as well as Christine McBride from the MACS school office.

August 2014
Edinburgh & Toulouse

Philippe Muller
Verena Rieser

Program Committee

Nicholas Asher	IRIT, CNRS
Timo Baumann	Universität Hamburg
Luciana Benotti	Universidad Nacional de Cordoba
Nate Blaylock	Nuance Communications
Holly Branigan	University of Edinburgh
Heriberto Cuayahuitl	Heriot-Watt University
Valeria De Paiva	University of Birmingham
Paul Dekker	ILLC, University of Amsterdam
David Devault	USC Institute for Creative Technologies
Arash Eshghi	Heriot-Watt University
Raquel Fernández	University of Amsterdam
Victor Ferreira	UC San Diego
Kallirroi Georgila	University of Southern California
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	King's College London
Markus Guhe	University of Edinburgh
Pat Healey	Queen Mary, University of London
Anna Hjalmarsson	Centre for Speech Technology, KTH
Amy Isard	University of Edinburgh
Simon Keizer	Heriot-Watt University
Ruth Kempson	Kings College London
Alexander Koller	University of Potsdam
Staffan Larsson	University of Gothenburg
Alex Lascarides	University of Edinburgh
Pierre Lison	University of Oslo
Colin Matheson	University of Edinburgh
Gregory Mills	University of Edinburgh
Philippe Muller (**chair**)	IRIT, Toulouse University
Chris Potts	Stanford University
Laurent Prévot	Laboratoire Parole et Langage
Matthew Purver	Queen Mary University of London
Hannes Rieser	Universität Bielefeld
Verena Rieser (**chair**)	Heriot Watt University, Edinburgh
David Schlangen	Bielefeld University
Gabriel Skantze	KTH
Amanda Stent	Yahoo! Labs
Matthew Stone	Rutgers
David Traum	ICT USC
Nigel Ward	University of Texas at El Paso

Contents

Poster Abstracts

Invited Talks

Holly Branigan
Professor for Psychology, University of Edinburgh

Say as I say: Alignment as a multi-componential phenomenon

Converging evidence from an ever-increasing number of experimental and observational studies suggests that people converge many aspects of their language (and other behaviour) when they interact. What is less clear is *why* such alignment occurs, and the function that it plays in communication. Discussions of individual instances of alignment have tended to appeal exclusively to one of three explanatory frameworks, focusing on social relationships between interacting agents, strategic maximisation of mutual understanding, or automatic linguistic priming behaviours. Each framework can satisfactorily explain some observed instances of alignment, but appears inadequate to explain others. I will argue that alignment behaviours are best characterised as multi-componential, such that all three kinds of mechanism may potentially and simultaneously contribute to the occurrence of alignment, with the precise contribution of each depending upon the context and aspect of language under observation. However, evidence from studies of typically developing children and speakers with Autistic Spectrum Disorder suggest that a tendency to align language may be in some sense 'wired in' at a very basic level, and that both the ability to suppress this reflex and the ability to strategically exploit alignment for social or communicative ends may be later acquired and superimposed on top of this basic and reflexive tendency.

Jon Oberlander
Professor of Epistemics in the University of Edinburgh

Talking to animals and talking to things

I will argue that to build the diverse dialogue systems that will help us interact with and through the Internet of Things, we need to draw inspiration from the dizzying variety of modes of human-animal interaction. The Internet of Things (IoT) has been defined as "the set of technologies, systems and methodologies that underpins the emerging new wave of internet-enabled applications based on physical objects and the environment seamlessly integrating into the information network". Although there is a technical view that the IoT will not require any explicit interaction from humans, it plausible to assume that we will in fact need to develop appropriate mechanisms to translate, visualise, access and control IoT data. We thus need to develop new means for humans to have 'words with things'. Some building blocks are already in place. Back in 2006, Bleecker proposed the 'blogject', an object that tracks and traces where it is and where it's been, has an embedded history of its encounters and experiences, and possesses some form of agency, with an assertive voice within the social web. In the last four years, this vision has been brought closer to reality through significant work on the "social web of things". But something is missing. The IoT will surely contain a huge variety of things, some with real intelligence and flexibility, and others with only minimal agency; some we will want to talk to directly; others will be too dull to hold a conversation with. Ever since Shneiderman's advice to the HCI community, we have struggled with the idea that if a system can sustain a multi-step dialogue, it must have human-level intelligence. So, in developing new ways to interact with the pervasive IoT, we must look beyond human-human interaction for models to guide our designs. Human-pet interaction is an obvious starting point, as in the work of Ljungblad and Holmquist, and recent projects on robot companions have already developed this line of thinking. However, pets represent just one point on the spectrum of human-animal interaction. Animals vary from wild, to feral, to farmed or caged, to working, through to domestic. Their roles include: companions (e.g. pets), providing aid and assistance (e.g. guide dogs), entertainment (e.g. performing dolphins), security (e.g. guard dogs), hunting (trained predators pursuing untrained prey), food (e.g. livestock), and scientific research participants (e.g. fruitflies). If we take into account the types and roles of the animals with which humans already interact, we can take advantage of existing understanding of the breadth of human-animal interaction, and evolve a rich ecosystem of human-thing dialogue systems.

Matthew Purver
Senior Lecturer, Cognitive Science Research Group, Queen Mary, University of London

Ask Not What Semantics Can Do For Dialogue - Ask What Dialogue Can Do For Semantics

Semantic frameworks and analyses are traditionally judged by sentential properties: e.g. truth conditions, compositionality, entailment. A semantics for dialogue must be consistent not only with these intrinsic properties of sentences, but with extrinsic properties: their distribution, appropriateness or update effects in context. The bad news, of course, is that this means our analyses and frameworks have to do more, and fulfilling these requirements has been the aim of a great deal of productive and influential research. But the good news is that it also means that dialogue can act as a "meaning observatory", providing us with observable data on what things mean and how people process that meaning -- data which we can use both to inform our analyses and to learn computational models. This talk will look at a few ways in which we can use aspects of dialogue --- phenomena such as self- and other-repair, situation descriptions, the presence and distribution of appropriate and informative responses --- to help us choose, learn or improve models of meaning representation and processing.

This talk describes joint work with a number of colleagues, but particularly Julian Hough, Arash Eshghi and Jonathan Ginzburg.

Michael Schober
Professor of Psychology, New School for Social Research

Dialogue, response quality and mode choice in iPhone surveys

As people increasingly communicate via mobile multimodal devices like iPhones, they are becoming accustomed to choosing and switching between different modes of interaction: speaking and texting, posting broadcast messages to multiple recipients on social media sites, etc. These changes in everyday communication practices create new territory for researchers interested in understanding the dynamics of dialogue. This talk will describe studies of 1200+ survey respondents answering survey questions from major US social surveys, either via voice vs. SMS text (native iPhone apps) and either with human vs. automated interviewers; because the studies contrast whether the interviewing agent is a person or automated and whether the medium of communication is voice or text, we can isolate effects of the agent and the medium. The studies measure completion rates, respondent satisfaction and response quality when respondents could and could not choose a preferred mode of responding; response quality was measured by examining "survey satisficing" (taking shortcuts when responding—providing estimated or rounded vs. precise numerical answers, and "straightlining"—providing the same responses to multiple questions in an undifferentiated way), reports of socially desirable and sensitive behaviors, and requests for clarification.
Turn-taking structure in text vs. voice is, of course, vastly different, with notably longer delays between turns in the asynchronous text modes, and greater reported multi-tasking while texting; and there were some notable differences in texting and talking with human vs. automated interviewers/interviewing systems. But the overall findings are extremely clear: notably greater disclosure of sensitive/embarrassing information in text vs. voice, independent of whether the interviewer is human or automated; and less estimation/rounding in text vs. voice, again independent of whether the interviewer is human or automated. The opportunity to choose a mode of interviewing led to improved satisfaction and improved response quality, with more respondents choosing text than voice. The findings suggest that people interviewed on mobile devices at a time and place that is convenient for them, even when they are multitasking, can give more trustworthy and accurate answers than those in more traditional spoken interviews. Survey interviews are a very particular kind of dialogue with particular constraints, but they are a useful laboratory for deeper understanding of the dynamics and pragmatics of dialogue.

Full Papers

Referring Expressions in Discourse about Haptic Line Graphs

Özge Alaçam
Department of Informatics
University of Hamburg
Hamburg/Germany
alacam@informatik.uni-hamburg.de

Cengiz Acartürk
Cognitive Science
Middle East Technical University, Ankara /Turkey
acarturk@metu.edu.tr

Christopher Habel
Department of Informatics
University of Hamburg
Hamburg/Germany
habel@informatik.uni-hamburg.de

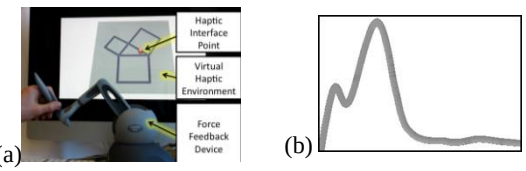
Abstract

Statistical line graphs are widely used in multimodal communication settings and they are crucial elements of learning environments. For visually impaired people, haptic-audio interfaces that provide perceptual access to graphical representations seem as an effective tool to fulfill these needs. In an experimental study, we investigated referring expressions used in a collaborative joint activity between haptic explorers of graphs and verbal assistants who helped haptic explorers conceptualize local and non-local second-order concepts (such as extreme values, trends, or changes of trends). The results show that haptic exploration movements evoke deictically referential links that are essential for establishing common ground between explorers and assistants.

1 Comprehending Graphs through Different Modalities

Data visualization aims at (re-)presenting data so that humans more easily access certain aspects of them (such as trends or anomalies) for thinking, problem solving and communication (Tufte 1983, Kosslyn 1989, 2006, Hegarty 2011, Alaçam, et al., 2013). Among many specific types of representational modalities (such as sketch maps, statistical graphs and schematic diagrams), statistical line graphs have found a widespread use in various daily life and professional settings. For making statistical graphs accessible to visually impaired people, technologies ranging from pure tactile graphs to verbal summaries (Demir et al., 2012) have been proposed. However, haptic presentations of graphs (henceforth, *haptic graphs*) provide a suitable means for visually impaired people to acquire knowledge from data sets, when they are integrated in hybrid systems that employ auxiliary modalities to the haptic-tactile modality, such as sonification and verbal assistance (Abu Doush et al., 2010; Ferres at al., 2013).

Users can explore haptic graphs by hand-controlling a stylus of a force-feedback device, for instance a Phantom Omni® (recently Geomagic® Touch™, see Figure 1.a), which yields information about geometrical properties of lines. Compared to visual graphs, one drawback of haptic graphs is the restriction of the haptic sense in simultaneous perception of spatially distributed information (Loomis et al, 1991). Comprehension of haptic line graphs is based on explorations processes, i.e. hand movements tracing lines, with the goal to detect shape properties of the graph line explored. The recognition of concavities and convexities, as well as of maxima and minima, is of major importance (see Figure 1.b for a sample haptic line graph).



(a) (b)

Figure 1. (a) Phantom Omni® device and visualization in a geometry domain (see, Kerzel & Habel, 2013, Fig. 1), (b) sample haptic graph

Although simple line graphs are often considered as a graph type easy to comprehend haptically, there are some critical problems about haptic representation of simple line graphs: Whereas it is only moderately difficult to comprehend the shape of a simple graph line with a single (global) maximum haptically, graphs with several local maxima require additional assistance for most users of haptic graphs. Providing additional information, such as aural assistance through the auditory channel, has been proved to be helpful for resolving some difficulties in haptic graph exploration (cf. sonification, Yu and Brewster, 2003). We propose to use speech utterances (i.e.

verbal assistance) to support—for example—the detection and specification of local and global extrema of graph lines, or other shape based concepts.

For designing haptic graph systems, which are augmented by computationally generated verbal assistance, it is necessary to determine which information, depicted by the graph or by its segments, are appreciated as important by haptic explorers. In this paper we focus on the use of referring expressions within dialogues in collaborative haptic-graph exploration-activities between blindfolded *haptic explorers* and seeing *verbal assistants*. The analyses of these joint activities provide crucial insight about how haptic explorers acquire high-level information from haptically perceived graphs. Moreover, they also provide the empirical basis (i.e. which spatial content should be verbalized) for our long-term goal: the realization of a cooperative system providing blind graph readers with verbal assistance (Habel et. al., 2013, Acartürk et. al, 2014).

1.1 Shape in Line Graphs: Perception, Cognition and Communication

Graph lines inherently convey shape information, namely information about convexities and concavities, about straightness, angles, and vertices. These are evoked in visual perception by visually salient graph-shape entities, in particular by curvature landmarks, positive maxima, negative minima, and inflections (Cohen & Singh, 2007).

From the perspective of a seeing human who describes a line graph, salient parts of the graph line are primary candidates to be referred to. In other words, referring expressions are evoked by visually salient graph entities. The conceptual inventory for verbalizing line-graph descriptions, as well as trend descriptions, has to fulfill requirements from language and perception. Since graph lines can be seen as a specific type of 2D-contours, we include some concepts proved as successful in visual shape segmentation into the inventory of spatial concepts, namely Cohen and Singh's curvature landmarks (2007). In addition to Cohen-Singh landmarks, the case of graph lines requires graph-line specific types of curvature landmarks: since graph lines are finite and not closed, two types of endpoints (left vs. right) have to be distinguished.

In haptic graph exploration the shape of the graph line is a major property for identifying referents by distinguishing it from its distractors. Additionally, certain aspects of graph segments (such as inflection points that show smooth change) are more difficult to acquire in the haptic modality than in the visual modality, largely due to the sequential and local perception with a narrow bandwidth of information in the haptic modality (Habel et. al., 2013). Finally, previous research has shown that not only saliency in the domain of discourse via the linguistic context but also saliency in the visual context influences humans' choice of referring expressions (Fukumura et al, 2010).

Haptic assistive systems that take shape properties of graphical representations into account in design process have been scarce except for a few instances (e.g. see Ferres et al., 2013; Wu et al., 2010). Additionally, there is still a lack of research on the role of shape comprehension in haptic graph exploration. Since the current state-of-the art haptic graph systems would benefit from providing verbal descriptions of shape properties and shape entities, we focus in this paper on the use of referring expression to these entities in collaborative graph explorations.

1.2 Assisted Haptic Graph Exploration: A Joint Activity Approach

Verbally assisted haptic graph exploration can be seen as a task-oriented collaborative activity between two partners, a (visually impaired) explorer (*E*) of a haptic graph and an observing assistant (*A*) providing verbal assistance (see Figure 2). Sebanz and colleagues (2006), who focus on bodily actions, describe joint actions as follows: "two or more individuals coordinate their actions in space and time to bring about change in the environment". In contrast to this characterization, the joint activities that we focus on shall bring about changes in *E*'s mental representations. To reach this goal, *E* and *A* have to establish common "understanding of what they are talking about" (Garrod & Pickering, 2004).

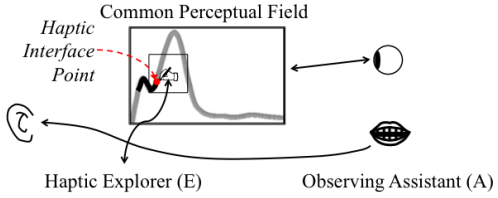


Figure 2. Assisted haptic graph exploration, a joint activity

A and *E* share a common field of perception, namely the haptic graph, but their perception and comprehension processes differ substantially. For example, while *E* explores the highlighted, black segment of the haptic graph, *A* perceives the global shape of the graph, in particular, *A* is

aware of shape landmarks and line segments. For example, when *E* explores the first local maximum followed by a local minimum (see Figure. 2), *E* does not have information about the global maximum, which is already part of *A*'s knowledge. Therefore, *E* and *A* have different internal representations of the graph line, and A's referring to the graph could augment *E*'s internal model substantially. For example, uttering "Now you have reached the heights of the last peak" would provide *E* with additional information. Another suitable comment would be "You are in the increase to the population maximum", or even "You are in the increase to the population maximum of about 90, that was reached in 1985". Since verbal assistance is a type of instruction, overspecified referring expressions are adequate for our domain (see Koolen et al., 2011).

The success of the joint activity of explorer *E* and observing assistant *A* in general, and also the success of *A*'s utterances in particular, depends, on the one hand, on joint attention (Sebanz, et al., 2006), and on the other hand, on the alignment of the interlocutor's internal models, especially on building implicit common ground (Garrod & Pickering, 2004). Since *E*'s internal model of the activity space, i.e. the haptic graph and E's explorations, is perceived via haptic and motor sensation, whereas *A*'s internal model of the same space is build up by visual perception, similarities and differences in their conceptualization play the central role in aligning on the situation-model level.

The assisted haptic graph explorations we discuss in this paper can be conceived as an asymmetric joint activity: firstly, the participants have different activity roles (explorer vs. assistant), as well as different sensor abilities; secondly, the participants were told that *E* should initiate the help request and *A* should provide help based on explorer's need. Although the dialogues accompanying haptic explorations are—in principle—mixed-initiative dialogues, explorer-initiatives are the standard case.

Haptic explorers' contributions to the dialogue are given concurrently to their exploration movements. Thus, for the observing assistant, the referring expressions produced are accompanied with the current exploration point on the graph. In other words, *E*'s exploration movement evokes deictically a referential link—analogue to Foster and colleagues' (2008) haptic ostensive reference. And thus, common ground is established and the given-new contract between *E* and *A* is fulfilled (Clark and Haviland, 1977; Clark and Brennan, 1991). In the following turn, *A* is expected to provide most helpful and relevant information for *E* at that particular moment. In particular *A* should provide *E* with content that is difficult to acquire haptically, such as, information about whether a maximum is local or global. To maintain the common ground, *A* has to synchronize her language production with *E*'s hand-movements in a turn-taking manner, since the quality of verbal assistance depends on establishing appropriate referential and co-referential links.

1.3 Shape Concepts in Graph-Line Descriptions

Most qualitative approaches to shape representation focus on the shape of contours (see, e.g., Hoffman & Richards, 1984; Eschenbach et al., 1998), and on curvature landmarks of contours (Cohen and Singh, 2007), such as, positive maxima and negative minima, depending on the concepts of convexity and concavity of contours, and inflection points. However, graph lines require some additional shape representations and shape cognition characteristics beyond the characteristics of contours. In particular, graph lines are conventionally oriented corresponding to reading and writing direction and they are comprehended with respect to an orthogonal system of two axes. The haptic graphs we use in the experiments are realized in a rectangular frame that induces an orthogonal system of axes. The geometric shape concepts for describing graph lines are exemplified with a graph used in our experimental studies (see Figure 3).

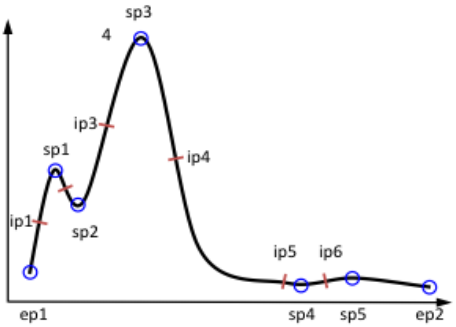


Figure. 3. Qualitative shape landmark ascription for a sample graph (augmented with orthogonal axes for making the reference frame in Table 1 explicit)

Table 1 gives a tabular summary of qualitative representations for selected shape landmarks and induced line segments. The functional character of statistical line graphs leads to the prominence of *value extrema* (in contrast to *curvature extre-*

ma of contours). Since we use in the experiments presented here *smoothed graphs*, these extrema are called *smooth points* (sp). *Inflection points* (ip) are depicted in Fig. 3 but not mentioned in Table. 1.)

Shape landmarks		
	Landmark characteristics	**Global properties**
ep1	left end pt., local min.	higher than sp4, ep2
sp1	smooth pt., local max.	higher than ep1, sp2, sp4, sp5, ep2
sp2	smooth pt., local min.	higher than ep1, sp4, sp5, ep2
sp3	smooth pt., local max.	global max.
sp4	smooth pt., local min.	same height as ep2
sp5	smooth pt., local max.	higher than sp4, ep2
ep2	right pt., local min.	same height as sp4
Shape segments		
	Shape characteristics	**Vertical orientation**
ep1–sp1	curved	steeply upward
sp1–sp2	curved	diagonally downward
sp2–sp3	curved	steeply upward
sp3–sp4	curved	steeply downward
sp4–sp5	curved	slightly upward
sp5–ep2	curved / nearly straight	slightly downward / nearly horizontal

Table 1. Qualitatively described shape landmarks and shape segments.

1.4 Referring to Shape Entities: Semantic Representations

Our long-term goal is to realize an automatic verbal assistance system that provides instantaneous support for haptic explorers during their course of exploration. Empirical studies are needed to study underlying principles of haptic graph exploration, and the effect of linguistically coded content in comprehension of second order entities, such as general and temporally restricted trends based on the recognition of global and local curvature landmarks.

The referring expressions produced by haptic explorers and verbal assistants during collaborative activity give insight about how graph readers comprehend graphs, which elements are mentioned most, and how they are referred to. The investigation of multimodal interactions (namely interaction by means of language, gesture and graph) requires systematic qualitative analysis, as well as quantitative analysis. We followed one of the widely accepted method developed by Dale and Reiter (1995), which addresses the generation of referring expressions, to characterize the semantic properties of graphical segments and the referring expressions produced during collaborative activity. In this paper, we do not aim to go into implementation level in detail, instead we used the method as a tool to make systematic mapping between semantic properties of graphical features and participants' referring expressions. According to Dale (1992), a system that generates referring expressions should at least satisfy Gricean-like conversational maxims targeting adequacy, efficiency and sensitivity. In more detail, a referring expression should contain enough information to allow the hearer to identify the referent, it should not contain unnecessary information and it should be sensitive to the needs and abilities of the hearer. They propose and implement a cost function that assumes (based on empirical research) people first and usually prefer to refer to type properties (zero cost), then to absolute properties. Relative properties and relations (the highest cost) follow them respectively. By following this method, we employed ⟨attribute, value⟩ pair representation to characterize the qualitative representations of graph shapes and landmarks. To illustrate, the attribute set which is available for the "*ep1-sp1*" shape segment (see Table 1) possesses the following properties: ⟨type, curved⟩, ⟨manner, steep⟩, and ⟨direction, up⟩. For the systematic data analyses, the verbal data produced in a joint activity were also characterized by using this method since it successfully foregrounds the common properties of multimodal data, see Table 2 for semantic attribute scheme for verbal data.

Type Properties:
- Terms
  - ⟨term, peak⟩, ⟨term, something⟩
- Location
  - Frame of Reference Terms ("start point")
  - Haptic Ostensive Expressions

Absolute Properties:
- ⟨value, 0⟩ for "it is 0"
- ⟨count, 3 peaks⟩

Relative Properties:
- ⟨size, small⟩ , ⟨manner, slowly⟩
- ⟨direction, up⟩

Relations:
- ⟨temporal relations, after the fall⟩
- ⟨spatial relations, higher⟩

Others:
- Interjections (hmm, ah…)
- Affirmations/Negations

Table 2. Semantic attribute scheme

In addition to the attributes stated by Dale and Reiter (1995), we identified haptic ostensive expressions (*HOE*s). The haptic explorers produced *HOE*s that referred to the pointed locations, which are also accompanied by assistance request from the verbal assistant. Foster and colleagues (2008) define the *HOE* as a reference,

which involves deictic reference to the referred object by manipulating it haptically. Since haptic explorer location is visible to verbal assistant during joint activity, haptic actions are useful to provide joint attention between E and A.

2 Experiment

2.1 Participants, Materials and Design

Thirty participants (fifteen pairs of sighted and blindfolded university students) participated in the experiment. The language of the experiment was Turkish, the native language of all participants. The experiment was conducted in single sessions and each session took approximately 1 hour (including warm-up & instruction sessions, exploration processes and post-exploration tasks). The sessions were audio/video recorded. Each participant pair was composed of a haptic explorer (*E*) and a verbal assistant (*A*). The participants were located in separate rooms so that they communicated through speakers without visual contact. During the experiment session, *E* explored the graph haptically and *A* was able to display the graph and the current location of *E*'s exploration, which was represented by an animated point marker on the visual graph presented at *A*'s screen. However, haptic pointing was possible only for E. The pairs explored informationally equivalent graphs, except for the difference in the modality of presentation (haptic and visual). Finally, *E* was instructed to explore the graph and ask for verbal assistance when needed by turning microphone on, whereas *A* was instructed to provide verbal assistance shortly and plainly, when requested by *E*. Before the experiment, a warm-up session was conducted to familiarize *E* with Phantom Omni® Haptic Device (Figure 1). After then, in the instruction session, the participants were informed that the graphs represented populations of bird species in a lagoon and also about post-exploration tasks detailed below. The graphs employed in this study were taken from a publicly available consensus report (*PRBO*, 2012). Each graph had a different pattern in terms of the number and polarity of curvature landmarks, length and direction of line segments. In the experiment session, each participant was presented five haptic line graphs in random order. Haptic graph exploration was performed by moving the stylus of the haptic device, which can be moved in all three spatial dimensions (with six degree-of-freedom). The haptic graph proper (i.e., the line of the line graph) was represented by engraved concavities on a horizontal plane; therefore haptic explorers perceived the line as deeper than the other regions of the haptic surface. The numerical labels were not represented. The participants did not have time limitation. After the experiment session, both participants (*E* and *A*) were asked independently to present single-sentence verbal descriptions of the graphs to a hypothetical audience. They also produced a sketch of the graph on paper. Two raters who are blind to the goals of the study scored the sketches for their similarity to the stimulus-graphs by using a 1 (least similar) to 5 (most similar) Likert Scale. The inter-rater reliability between the raters was assessed using a two-way mixed, consistency average-measures *ICC* (Intra-class correlation). The resulting *ICC* (*=.62*) was in the "good range" (Cicchetti, 1994).

3 Results

The participants produced 75 dialogues (5 stimuli x 15 pairs). The data from two pairs were excluded since they did not follow the instructions. The remaining 65 dialogues were included into the analysis. The average length of a dialog was 103 seconds (*SD*=62 sec.). The results of this experiment, which focus on the role of taking initiative for assistance, were reported elsewhere (Alaçam et. al. 2014). In the present study, we focus on the semantic representation method and the production of haptic ostensive expressions during joint activity. Each utterance in the dialogues was transcribed and time-coded. The transcriptions were then annotated by the semantic attribute scheme presented in Table 2. The term "utterance" refers to speech parts produced coherently and individually by each participant. We classified the utterances into three categories; (i) Request-Response Pairs, (ii) Alerts initiated by *A* (but do not require response from *E*) and (iii) think-aloud sentences. In total, 1214 individual utterances were produced by the participants. 449 of them were initiated by the haptic explorers to communicate with their partners, 402 of them were produced by the verbal assistants as a reply to *E*. Those two types comprise 70.1% of all utterances. 65 utterances (5.35%) were initiated by *A*s. Utterances that were initiated by *A*s, without a request from *E* were mostly the utterances that alerted *E* when s/he reached to a start point or an end point. Although *E*s were not instructed to use the think-aloud protocol, self-talking during haptic exploration was observed in 10 of 13 haptic explorers. Those think-aloud sentences (i.e. the sentences without a communica-

tion goal with the partner since the explorers did not turn on microphone during self-talking) constituted 24.5% of all utterances (*N*=298). In this paper we focused on the communicative utterances, therefore we restricted our analysis to "Request-Response Pairs" and "Alerts" excluding "Think-aloud" sentences. The results pointed out that the most frequently observed assistance content was about information for positioning, such as being on a start point or end point, on the frame, or being inside or outside of the line. 72.4% of the utterances (341 utterances in total - 46 of them initiated by *A*) addressed this type of information.

*E*s showed a tendency to request assistance by directing "Yes/No Questions or Statements" to *A*s (*N*=418) instead of using open-ended questions (*N*=7). *A*'s contributions to the dialogue can be also classified as follows: (1) instructional, *N*=69 (i.e. navigational, such as 'go downward from there'), or (2) descriptional utterances, *N*=386. Descriptional utterances included, (2a) confirmative assistance, *N*= 342 (confirming the information which haptic explorer has already access), and (2b) additional assistance, *N*=44 (introducing new property or updating the value of already stated property). Below we present sample request-response pairs, which introduced new information or updated the value of the already introduced attribute.

- *E*: Is this the start point? *A*: Yes, it is also the origin (*A* updates ⟨type, start point⟩ as ⟨type, origin⟩ that emphasizes 2D frame of reference, and that implicitly carries over the value for the starting point)
- *E*: no request. *A*: You are at the first curve; ⟨type, curve⟩, ⟨relation, order, first⟩ (both type and relation attributes were introduced to the dialogue)

The non-parametric correlation analyses using Kendall's tau showed positive correlation between the existence of attribute update in the dialogue and higher sketching scores (*N*=62, τ=.46, *p*=<.01). Moreover, the number of attribute updates is positively correlated with higher sketching scores (*N*=62, τ=.45, *p*=<.01). As an illustration, consider one of the dialogues between *E* and *A*: *E* asked a question ("*Is this going perpendicular?*") to *A* by pointing "*ep1-sp1*" segment of the graph presented in Figure 3. As stated in Table 1, this shape segment can be labeled with ⟨type, curved⟩, ⟨manner, steep⟩, ⟨direction, up⟩ attributes. In his question, *E* addresses both manner and direction attributes. However, the word for "perpendicular" in Turkish can be used to refer to both being perpendicular and steep. Here *A´*s response ("*There is a slight slope*") updates *E´*s information and it also clarifies possible misunderstanding, since in statistical graphs in time domain, perpendicular lines are not allowed. The resulting request-response pair covers all attribute pairs for the particular graph shape (the region which *E* needs assistance) and the sketch was rated with 4.5 in average (in 1to5 Likert Scale). The parameters (Dale and Reiter, 1995) (i) the number of attributes that are available to be used in a referring expression and (ii) the number of attributes mentioned in the final expressions seem as a useful indicator to evaluate the successful communication.

Additionally, verbal assistants' expressions that referred to a point or a region on the graph, namely type property, were mostly graph-domain terms (such as "curve", "peak" etc.). On the other hand, haptic explorers showed a tendency to use simpler expressions such as "something", "hill", "elevation". This indicated that haptic explorers had difficulty to access graph-domain vocabulary to name the regions or the shape, so that they choose alternative ways to name it (including use of onomatopoeic words such as "hop hop").

The haptic ostensive actions and expressions performed to catch the attention of the assistant do not directly contribute to conceptualizing the graph shape; still their communicative role in the dialogues is important. 20.4% (*N*=247) of all the communicative utterances contained *HOE* that enhanced the reference resolution, therefore shorter descriptions could be produced instead of long descriptions. The analysis of verbal data revealed two major subcategories of *HOE*s: (i) Demonstrative Pronouns (*DP*s) such as "This/Here" or "like this"), and (ii) temporal pointings (*TP*s) such as "Now". Table 3 illustrates the frequency values for each *HOE* category. Non-parametric Wilcoxon Signed-Rank tests were conducted to investigate the use of different *HOE* types. The results showed that the haptic explorers produced more *DP*s (*z*=-4.88, *p*=<.001) and *TP*s (*z*=-3.75, *p*=<.001) than the assistants produced. While there is no significant difference in the number of *DP*s and *TP*s produced by *E*s (*z*=-.50, *p*=>.05), *A*s preferred to use *TP*s rather than *DP*s. Only a few instances (*N*=5) of *DP*s uttered by *E* was responded by *A*'s use of *DP*s. The instances that illustrate *A´*s responding to *E* by using different *HOE* category than the one used by *E* were not observed at all.

	Only by *E*	Only by *A*	Both *E* & *A*
Demonstrative Pronoun-*DP*	99	6	5
Temporal Pointing-*TP*	67	27	19

Table 3. The number of *HOE*s for each category

We performed a further analysis on salient graph parts by focusing on in which area of the graph the participants preferred to use one of the two *HOE* categories (demonstratives and temporal pointing) for referring. For this, the accompanying content (location being referred to) were classified into three groups, (i) reference to start points and end points, (ii) reference to intermediate points or regions on the graph and (iii) reference to frame (such as being on the frame, or being outside of the line). The results of the analysis showed a significant association between the referred location and the *HOE* preference, $X^2(2)=38.2$, $p<.001$. The results (the standard residuals for each combination) indicated that when the participants referred to a start/end point of the graph line, they used *DP*s ($N=48$, $z=-.6$) and *TP*s ($N=48$, $z=-.7$). However, for referring to any particular point or any region on the graph, they preferred *DP*s ($N=59$, $z=2.8$) rather than *TP*s ($N=16$, $z=-3.1$). Moreover, when they mentioned about the events related to the reference frame, they preferred *TP*s ($N=29$, $z=3.3$) rather than DPs ($N=6$, $z=-3$, *all p values are smaller than .05*). However no main association was found between *HOE* types (*DP*s or *TP*s) and whether the referred region is a point or area. This indicates that both specific points (i.e. landmarks) and broader regions (i.e. line segments) haptically highlighted by *E* were accompanied by any of *HOE* types; however the position of the point or region on the graph (i.e. at the beginning or at the intermediate region on the line) has effect on which *HOE* type is preferred.

4 Discussion

In an experimental setting, which employed a joint-activity framework, pairs of participants (haptic explorers and verbal assistants) explored the graphs and they exchanged verbal information when necessary. Following Dale and Reiter (1995), we categorized graph shapes (segments/landmarks) and verbal data as attribute pairs such as ⟨type, maximum⟩. When *E* needs assistance about a segment, or global shape, her/his question was modeled as a specification of the choices of some of the attributes. As a response to the request for assistance, the description of *E* may be complete, lacking or partially or completely inaccurate. In order to have successful communication, verbal assistant should provide lacking information or correct the incorrect interpretation to complete the coverage of attributes in "target set" of attributes. Within this framework then, we assume that successful communication is achieved when *E* requests assistance (initiated by haptic explorer w.r.t. his needs to avoid over-assistance) and *A* updates the attribute pairs or introduces new attributes. Moreover, since *E* already has access to basic spatial properties, a useful solution would be to provide information with graph-domain terms, and relative terms (since absolute terms are difficult to implement), as well as relational terms that emphasize size and manner gradually (w.r.t. haptic explorer´s needs and current knowledge). The results of the experiment also showed that *A*'s role in *E*´s comprehension is critical. First, *A* has a more complete mental representation of the graph starting from the onset of haptic exploration due to spontaneous visual exposure to both global and local information on the graph. Their guidance on salient points with additional attributes or their aligning the instructions w.r.t haptic explorer´s current understanding of the graph enhances the comprehension of *E*. Moreover, the verbal assistants introduced more graph domain oriented concepts to dialogues, while haptic explorers tended to use simpler daily terms or even onomatopoeic words. This information is important when forming attribute set for graph shapes.

Our focus was to investigate the content that needs additional assistance but our results also pointed out the information that can be provided more effectively by a different modality than verbal modality. The research by Moll and Sallnäs (2009) and Huang et al (2012) suggest audio-haptic guidance for visually impaired people to enhance navigational guidance in virtual environments so that the participants focus on communication at a higher level. Their results indicated that "by using haptic guiding one can communicate information about direction that does not need to be verbalized" (Moll and Sallnäs, 2009, p.9) and "sound provides information that otherwise has to be conveyed through verbal guidance and communication" (Huang et al., 2012, p.265). Considering that 72.4% of the utterances in our experiment contained information about positioning (being on the start point, or on the line etc.), providing this information to the explorer seems crucial for the assistive system; however delivering this infor-

mation verbally would yield continuously speaking assistance, therefore sonification can be a good candidate to carry this message. Additionally, haptic exploration allows haptic ostensive actions that highlight the attended location. The location attribute has different characteristics than other attribute pairs. It grounds joint attention between partners by pointing where the assistance is needed, then other attributes provide additional information about what the graph shape means. As for *HOE*s, the type of referring expressions (demonstrative pronouns or temporal pointing) seems affected by the referred location (start/end points, intermediate regions or graph frame). The results also indicated that the explorers produce significantly more *HOE*s during joint activity compared to the verbal assistants. In the collaborative activity settings that allow both users (the human explorer/learner and human or robot assistant) to manipulate the environment haptically (Foster et al., 2008; Moll and Sallnäs, 2009), the assistants' haptic ostensive actions have salient communicative function. However, in our assistance setting, only haptic explorers have active role in the haptic exploration. Even after requesting assistance from A regarding specific point or region by pointing with *HOE, E* may still continue to explore. Therefore verbal assistants tend to omit uttering *HOE* and when necessary, they use temporal indicators to relate a previously mentioned expression to currently explored region. This preference of verbal assistants may be due to prevent explorers' incorrect reference resolution.

Finally, in addition to attribute-set approach of Dale and Reiter (1995), a more context sensitive version that implemented salience weights was proposed by Krahmer and Theune (2002). The comparative study between visual and haptic perception of graphs indicated that haptic readers tend to overestimate small variations on the graph shape due to haptic salience induced by haptic friction and to underestimate smooth regions that can be useful for segmentation (Habel et. al, 2013). Choosing appropriate attribute value enhanced with salience weights for this kind of haptically problematic regions might overcome this problem in the implementation level.

5 Conclusion

Graphs are one of the efficient ways of visual communication to convey the highlights of data, however visual perception differs from haptic perception; therefore the highlighted piece of information in visual modality can be hidden when it is converted to haptic modality. Hence, investigation of differences in two modalities is necessary to detect and close the informational gap. The current study that explores on-line haptic graph comprehension in the presence of verbal assistance contributes our understanding about haptic graph comprehension by investigating dialogues between haptic explorer and verbal assistant as a collaborative activity.

Taking the Gricean Maxims into account in the generation of referring expressions (careful selection of the information provided in "attribute pairs", updating attributes gradually and being sure that at the end of the communication target attribute set is covered) seems useful in enhancing the conversational success of the communication (Grice, 1975; Dale, 1992; Dale & Reiter, 1995). In contrast to providing all likely information to the graph reader all together, the detection of what s/he wants to know at a particular time would yield a more effective design of the (learning) environment for the graph reader when we take into account his/her current position, previous haptic exploration movements and utterances (the referred locations and how these regions were referred), thus addressing adequacy, efficiency and sensitivity criteria. For this reason, semantic mapping needs to be accomplished in multimodal data. Following Dale and Reiter´s approach, we represented graph shapes and verbal data as attribute pairs in the present study. The empirical results revealed that a more successful communication was observed when the attributes used by haptic explorers were enriched by means of specific, graph-domain terminology. Accordingly, building up a multimodal system based upon this approach looks promising. Future work will address designing the generation of verbal assistance based on the experimental findings.

Acknowledgments

The study reported in this paper has been supported by DFG (German Science Foundation) in ITRG 1247 'Cross-modal Interaction in Natural and Artificial Cognitive Systems' (CINACS) and by METU BAP-08-11-2012-121 'The Study of Cognitive Processes in Multimodal Communication'.

References

Abu Doush, I., Pontelli, E., Simon, D., Son, T.C., & Ma, O. (2010). Multimodal Presentation of Two-Dimensional Charts: An Investigation Using Open Office XML and Microsoft Excel. *ACM Transactions on Accessible Computing, 3*, 8:1–8:50.

Acartürk, C., Alaçam, Ö., & Habel, C. (2014). Developing a Verbal Assistance System for Line Graph Comprehension. In A. Marcus (Ed.): *Design, User Experience and Usability* (DUXU/HCII 2014), Part II, (pp. 373–382). Berlin: Springer-Verlag.

Alaçam, Ö., Habel, C. & Acatürk, C. (2014). Verbally Assisted Haptic Graph Comprehension: The Role of Taking Initiative in a Joint Activity. To be published in the Proceedings from the 2st European Symposium on Multimodal Communication, University of Tartu, Estonia, August 6-8, 2014.

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290.

Clark, H., & Haviland, S. (1977). Comprehension and the Given-New Contract. In: R. O. Freedle (ed.), *Discourse Production and Comprehension* (pp. 1–40). Erlbaum, Hillsdale, NJ.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared* cognition (pp. 127-149). American Psychological Association, Washington, DC.

Cohen, E., & Singh, M. (2007). Geometric Determinants of Shape Segmentation: Tests Using Segment Identification. *Vision Research, 47*, 2825-2840.

Dale, R. (1992). *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, MA.

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

Demir, S., Carberry, S., & McCoy, K.F. (2012). Summarizing Information Graphics Textually. *Computational Linguistics, 38*, 527–574.

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

Ferres, L., Lindgaard, G., Sumegi, L., & Tsuji, B. (2013). Evaluating a tool for improving accessibility to charts and graphs. *ACM Transactions on Computer-Human Interaction ,20(5)*, 28:1–28:32.

Foster, M.E., Bard, E.G., Hill, R.L., Guhe, M., Oberlander, J., & Knoll, A. (2008). The Roles Of Haptic-Ostensive Referring Expressions in Cooperative, Task-based Human-Robot Dialogue. In *Proceedings of ohe 3rd ACM/IEEE International Conference on Human-Robot Interaction*, pp. 295-302. Amsterdam, March 12-15, 2008.

Fukumura, K., van Gompel, R., & Pickering, M. J. (2010). The Use of Visual Context During the Production of Referring Expressions. *Quarterly Journal of Experimental Psychology 63*, 1700–1715.

Garrod, S., & Pickering, M. J. (2004). Why is Conversation so Easy? *Trends in Cognitive Sciences, 8*, 8-11.

Grice, H.P. (1975). Logic and conversation. In P.Cole & J. Morgan (Eds.), *Syntax and Semantics: Vol 3, Speech acts* (pp.43-58). New York: Academic.

Habel, C., Alaçam, Ö., & Acartürk, C. (2013). Verbally assisted comprehension of haptic line-graphs: referring expressions in a collaborative activity. *In Proceedings of the CogSci 2013 Workshop on Production of Referring Expressions*, Berlin.

Hegarty, M. (2011). The Cognitive Science of Visual-spatial Displays: Implications for Design. *Topics in Cognitive Science, 3*, 446–474.

Hoffman, D. & Richards, W. (1984). Parts of recognition. *Cognition, 18*, 65–96.

Huang, Y. Y., Moll, J., Sallnäs, E. L., & Sundblad, Y. (2012). Auditory Feedback in Haptic Collaborative Interfaces. *International Journal of Human-Computer Studies, 70(4)*, 257-270.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors Causing Overspecification in Definite Descriptions. *Journal of Pragmatics, 43*, 3231-3250.

Krahmer, E., & Theune, M. (2002). Efficient Context-sensitive Generation of Referring Expressions. In: K. van Deemter & R. Kibble (Eds.) *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. (pp. 223-264). CSLI Publications, Stanford.

Kerzel, M. & Habel, C. (2013). Event Recognition During Exploration of Hapic Virtual Environment Line-based Graphics. In T. Tenbrink, J. Stell, A. Galton & Z. Wood (eds.) *Spatial Information Theory, 11th International Conference, COSIT 2013*. (pp. 109–128). Berlin: Springer-Verlag.

Loomis, J., Klatzky, R., & Lederman, S. (1991). Similarity of Tactual and Visual Picture Recognition with Limited Field of View. *Perception*, 20, 167-177.

Moll, J., & Sallnäs, E. L. (2009). Communicative Functions of Haptic Feedback. In: M. E. Altinsoy,

U. Jekosch, & S. A. Brewster (Eds.), *Haptic and Audio Interaction Design*. (pp. 1-10). Springer, Berlin Heidelberg.

PRBO. Waterbird Census at Bolinas Lagoon, MarinCounty, CA. Public report by Wetlands Ecology Division, Point Reyes Bird Observatory (PRBO) Conservation Science. (2012) http://www.prbo.org/cms/366, retrieved on January 29, 2012.

Sebanz, N, Bekkering, H., & Knoblich, G. (2006). Joint Action: Bodies and Minds Moving Together. *Trends in Cognitive Sciences, 10*, 70-76.

Spanger, P., Yasuhara, M., Iida, R., Tokunaga, T., Terai, A., & Kuriyama, N. (2012). REX-J: Japanese Referring Expression Corpus of Situated Dialogs. *Language Resources and Evaluation, 46*, 461-491.

Wu, P., Carberry, S., Elzer, S., & Chester, D. (2010). Recognizing the Intended Message of Line Graphs. In: Goel, A.K., Jamnik, M., & Narayanan, N.H. (eds.) *Diagrammatic Representation and Inference*. (pp. 220–234). Springer, Heidelberg.

Yu, W., & Brewster, S.A. (2003). Evaluation of Multimodal Graphs for Blind People. *Journal of Universal Access in the Information Society 2*, 105-124.

A Dynamic Minimal Model of the Listener for Feedback-based Dialogue Coordination

Hendrik Buschmeier and **Stefan Kopp**
Social Cognitive Systems Group — CITEC and Faculty of Technology
Bielefeld University, Bielefeld, Germany
{hbuschme, skopp}@uni-bielefeld.de

Abstract

Although the notion of grounding in dialogue is widely acknowledged, the exact nature of the representations of common ground and its specific role in language processing are topics of ongoing debate. Proposals range from rich, explicit representations of common ground in the minds of speakers (Clark, 1996) to implicit representations, or even none at all (Pickering and Garrod, 2004). We argue that a minimal model of mentalising that tracks the interlocutor's state in terms of general states of perception, understanding, acceptance and agreement, and is continuously updated based on communicative listener feedback, is a viable and practical concept for the purpose of building conversational agents. We present such a model based on a dynamic Bayesian network that takes listener feedback and dialogue context into account, and whose temporal dynamics are modelled with respect to discourse structure. The potential benefit of this approach is discussed with two applications: generation of feedback elicitation cues, and anticipatory adaptation.

1 Introduction

Communicative feedback (*mhm, okay,* nodding, and so on) is a dialogue coordination device used by listeners to express their mental state of listening — e.g., I understand what you say (Allwood et al., 1992) — and by speakers to hypothesise about this mental state and adapt their language production accordingly — e.g., she understood it, I can provide new information (Clark and Krych, 2004). One crucial question from the speaker's perspective is how listener feedback signals can be interpreted in the dialogue context, and how they relate to what has been or is being said. Listeners can, in principle, produce feedback signals at any point of time in a dialogue — without having to take the turn. There is also no restriction on the number of feedback signals that can be placed within a dialogue segment, whether it is a turn, an utterance, a pause or a combination of these. Consider the dialogue in example (1):

(1) KDS-1, U01 (9:46–9:58)[1]

```
1 S1: genau
2     allerdings ist Badminton da=
      =wieder verschoben
3     [weiß nicht] ob das jetzt=
  U1: [mhm       ]
  S1  =dauerhaft ist (.)
4 S1: [aber die zwei] Wochen=
  U1: [okay         ]
5 S1: =hab ich's jetzt so drin
  U1:                        ja
6 S1: das is wieder von=
7     =ehm acht bis zweiundzwanzig
      U[hr]
  U1:  [ok]ay (0.34)
8     ja,
9     dann ehm geh ich da trotzdem=
      =hin (.) ...
```

Speaker S1 explains to her interlocutor U1 that the regular badminton training has (again) been moved to a different time, and now takes place from 8 to 10 p.m. She also says that she does not know whether this change is permanent, but that it is scheduled like this for the next two weeks. During S1's nine seconds short turn (1.1) to (1.7), U1 provides four instances of communicative feedback. Firstly, she signals understanding with *mhm*, simultaneously producing a single head nod and looking at S1 (1.3). After that, she signals acceptance of the speaker's ignorance concerning the permanency of the time change with an *okay* that is accompanied by

[1] Excerpt from the calendar assistant domain corpus KDS-1 (http://purl.org/scs/KDS-1). Overlapping talk is marked with aligned square brackets. The transcription follows the GAT 2 system (Couper-Kuhlen et al., 2011).

a head nod (1.4). Thirdly, she signals understanding, producing a short and prosodically flat *ja*, German for 'yeah', (1.5). And finally, with S1 gazing at her, she signals understanding of the new time with an *okay* and a head nod (1.7). After a pause, U1 then takes the turn and continues.

In previous work (Buschmeier and Kopp, 2012), we proposed a Bayesian network approach in which single instances of communicative feedback are interpreted in terms of a few general attributes (contact, perception, understanding, acceptance, and agreement; Allwood et al., 1992). However, when multiple feedback instances occur in sequence, as in the dialogue in example (1), the question arises how their interpretations affect each other, and how they relate to what has been and is being said. In keeping with this 'minimal mentalising' approach to the listener's cognitive state, we take the Bayesian network model and make it dynamic. The dynamics is added by extending the model with a temporal dimension that accounts for the incremental and dynamic nature of dialogue. Thus, in this work, we propose a 'dynamic minimal model' of mentalising which can naturally deal with multiple instances of feedback by updating its representation — taking the immediate dialogue history into account as well — when the dialogue proceeds and feedback occurs.

2 Common ground and feedback

Participating in dialogue involves more than utterance planning, formulation, speaking, listening and understanding. One central task for interlocutors is to track the 'dialogue information state,' a rich representation of the dialogue context. The representation includes which information is grounded and which is still pending to be grounded; which knowledge is private and which is believed to be shared; who said what, how and when; how these utterances are related to each other; which objects have been introduced and are accessible for anaphoric reference; what is the current question under discussion; who is having the turn; and potentially much more (Clark, 1996; Larsson and Traum, 2000; Asher and Lascarides, 2003; Ginzburg, 2012).

In general, maintaining (i.e., representing and constantly updating) an information state is thought to be crucial for being able to successfully participate in dialogue. The necessity of some parts, such as a representation of accessible referents, is agreed upon among researchers. Without this information being maintained, typical dialogues would simply not be possible. Concerning the representation of common ground, however, researchers do not agree on how deep and rich it needs to be and how exactly it is used in language production.

On the one hand, Clark (1996) argues that interlocutors maintain a detailed model of common ground, even to the extent that mutual knowledge (approximated with various heuristics) is necessary to explain certain phenomena in language use (Clark and Marshall, 1981). Pickering and Garrod (2004), on the other hand, believe that dialogue does not involve heavy inference on common ground at all, instead they claim that primed and activated linguistic representations provide sufficient information in themselves.

Use of common ground in language production in dialogue is also a topic of ongoing debate. Clark (1996) and Brennan and Clark (1996) argue that common ground is critical in collaborative discourse. Utterances are designed in such a way that common ground as well as shared knowledge are taken into account. Since this might be cognitively too demanding, Galati and Brennan (2010) propose a lightweight 'one-bit' partner model (e.g., whether the addressee has heard something before or not) that can be used instead of information about full common ground and shared knowledge when producing an utterance. Horton and Keysar (1996) go even further and present evidence that language production is, at its basis, an egocentric process — interlocutors do not take common ground into account when initially planning an utterance unless they identify a possible problem while monitoring utterance execution. Finally, Pickering and Garrod (2004) claim that the only factors guiding language production are priming, activation, and, if necessary, interactive repair.

Speakers infer groundedness and common ground based upon 'evidence of understanding' of the interlocutors (Clark, 1996). One way for listeners to show such evidence is by providing communicative listener feedback as, e.g., short verbal/vocal expressions such as *mhm, okay,* and *oh;* head-gestures such as nods or shakes; facial expressions such as surprise, or frowning; as well as various gaze behaviours. Listener feedback is a particularly interesting kind of evidence of understanding for multiple reasons:

1. When providing feedback, listeners do not need to have or to take the turn, making it

very *fast*. Since it is not constrained by turn-taking, feedback can be given as soon as the need arises, enabling speakers to quickly adapt the ongoing utterance based on this information.

2. At the same time, feedback is *unobtrusive* and does not interrupt speakers during their utterance. It happens in the 'back channel' of communication (Yngve, 1970). Feedback also relies heavily on non-verbal modalities (head, face, gaze) that do not interfere with the speakers' linguistic processing. Verbal/vocal feedback expressions — that have the potential to interfere — are often non-lexical (Ward, 2006), usually short, and even prosodically hidden in the speech context provided by the speaker (Heldner et al., 2010).

3. Despite their shortness, feedback signals are very *expressive*. They are rich in their form (Ward, 2006) — enabling a fine-grained expression of subtle differences in meaning —, multi-functional, and interact heavily with their dialogue context (Allwood et al., 1992). Feedback is only partially conventionalised, relying on iconic properties instead.

4. Finally, communicative feedback is *reflective* of the listener's cognitive state with respect to language and dialogue processing. It indicates (or is used to signal) whether listeners are in contact with speakers, whether they are able and willing to perceive or understand what is being or has been said, whether they are able and willing to accept the message and what their attitude is towards it (Allwood et al., 1992). Furthermore, depending on its prosodic realisation, its placement, or its timing, feedback may also be indicative of the listeners' uncertainty about their own mental state, their urgency for providing feedback, the importance of this feedback item, and more such qualifiers to its basic communicative functions (Petukhova and Bunt, 2010).

Because of these properties, listener feedback is a viable basis for estimating groundedness and common ground. Since the communicative functions of listener feedback reflect the interlocutor's internal state, a somewhat detailed picture of the interlocutor (and hence the dialogue) can be formed based on it. Especially the latter two properties suggest that

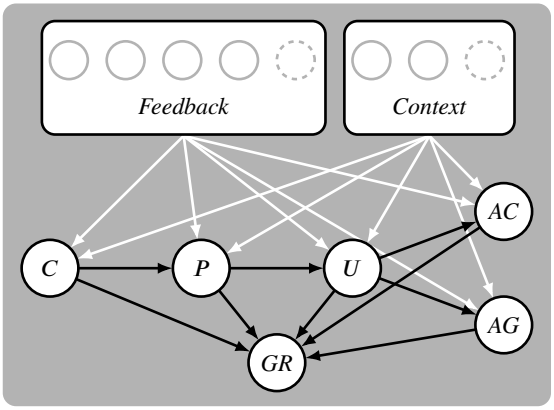


Figure 1: The Bayesian network model of the 'attributed listener state' (ALS; Buschmeier and Kopp, 2012). The random variables C, P, U, AC, and AG model a speaker's degree of belief that a listener is in contact, whether he or she perceives, understands, accepts, and agrees to what is communicated. A speaker's belief in groundedness is informed by all five of these variables.

feedback facilitates a form of mentalising about the cognitive state of the dialogue partner that goes beyond what is usually considered groundedness.

In previous work (Buschmeier and Kopp, 2012), we modelled this capability of speakers as, what we called, an 'attributed listener state' (ALS, cf. Figure 1). The ALS is a Bayesian network-based representation of a speaker's belief of what her listener's cognitive state is in terms of the basic communicative functions underlying feedback in dialogue. Each of the random variables (i.e., the nodes of the network) represent one 'dimension' of the multidimensional cognitive state of the listener: C (is the listener believed to be in contact), P (is the listener believed to perceive), U (is the listener believed to understand), AC (is the listener believed to accept), and AG (is the listener believed to agree). The network captures the dependencies between these variables and models their interactions, e.g., their hierarchical properties (Allwood et al., 1992; Clark, 1996). A belief about the groundedness of the conveyed proposition is formed based on the five ALS-variables, each having a different strength of influence.

The variables consist of the individual elements *low*, *medium*, and *high*, denoting whether the speaker believes the dimension of a listener's cognitive state to be low, medium, or high, respectively. An individual element's probability, e.g., $P(U = \mathit{low}) = 0.6$, is thus interpreted as the speaker's de-

gree of belief in this dimension of the listener's cognitive state to have the specific characteristic, i.e., 'with a probability of 0.6 the listener's understanding is believed to be low'. The probability distribution over all elements of a variable represents the speaker's belief state over the variable.

Buschmeier and Kopp's (2012) model can be considered a *minimal* form of mentalising based on listener feedback. It shares some desirable properties with the lightweight 'one-bit' partner model of Galati and Brennan (2010) — efficient processing in contrast to models of full common ground, a simple variable-based representation — while extending it. In particular, the model is in accordance with gradient representations of common ground (Brown-Schmidt, 2012), as it defines groundedness of a segment on an ordinal, non-binary scale (low < medium < high). Due to its probabilistic nature, each element is associated with a degree of belief from 0 (not believed) to 1 (believed). This information can be used to interactively adapt language production to a listener's need, e.g., by repeating/leaving out parts of an utterance, by giving subsequent parts a lower/higher information density, or by making information pragmatically explicit/implicit (Buschmeier et al., 2012).

3 A dynamic model of the listener

What is missing from the model proposed by Buschmeier and Kopp (2012), however, is a notion of the temporal dynamics that would make the evolution of the ALS coherent and continuous, and enable the model to deal with sequences of feedback such as in the example dialogue (1).

We regard an unfolding dialogue as a sequence of segments $[s_{t_0}, s_{t_1}, \dots s_{t_n}]$, each consisting of a dialogue move of the speaker (Poesio and Traum, 1997), together with any feedback responses of the listener. The static model of Figure 1 (Buschmeier and Kopp, 2012) treats each of these segments s_{t_i} independently and thus only reasons about the listener's cognitive state during one single segment. When doing the listener state attribution for the next segment, information from the preceding segments is not taken into account at all. To overcome this limitation, i.e., to account for the evolution of the listener's cognitive state over time, we need to give the model of the listener a temporal dimension.

As Bayesian networks are, in general, not limited in the number of edges and nodes, it would be possible to capture a whole dialogue — or at least a self contained and coherent part of a dialogue — in one large network that consists of connected sub-networks ALS_{t_i} — each corresponding to the network in Figure 1 — one for each segment s_{t_i}. The variables in the sub-networks would be uniquely named, and the networks evidence variables would be instantiated from the listener's feedback behaviour as well as the dialogue context of segment s_{t_i}. Furthermore, the variables between the sub-networks could be arbitrarily connected to model any desirable interaction between feedback and context across segments.

Theoretically, this approach could even work in an incremental framework. With each new dialogue segment $s_{i_{t+1}}$, a new sub-network $ALS_{t_{i+1}}$ would be added and connected to the network and Bayesian network inference would be carried out. However, even though there is, in principle, no limit in the size of a Bayesian network, the computational costs are rising polynomially with the number of nodes, and may even become intractable if the nodes are unfavourably connected (Barber, 2012). This makes this 'growing network approach' unsuitable for practical applications.

A slightly more constrained approach is to make a first-order Markov assumption, i.e., to assume that variables $X_{t_{i+1}}$ of a sub-network $ALS_{t_{i+1}}$ are only dependent on variables X_{t_i} of the sub-network ALS_{t_i} that directly precedes it. This can be achieved efficiently in the framework of *dynamic Bayesian networks*. In contrast to a constantly growing network approach, the dynamic Bayesian network approach consists of a maximum of two sub-networks ('time-slices') at any point of time. In such a *two time-slice Bayesian network* (cf. Figure 2), one time slice ALS_{t_i} represents the current dialogue segment s_{t_i} the other time slice the next segment s_{i+1}. As in the growing network approach, temporal influences among dialogue units are modelled by connecting some of the variables between the time-slices. Connection further back are, however, not possible.

In such a network, evolution over time is done by unrolling the network. Bayesian network inference is carried out on time-slice ALS_{t_i} and the resulting marginal posterior probabilities of those variables X_{t_i} that have a connection with variables $X_{t_{i+1}}$ in the next time-slice are computed. These posteriors are then used as 'prior feedback' (Robert, 1993), i.e., they are interpreted as prior distributions of those variables X_{t_i} that are used as evidence variables to variables $X_{t_{i+1}}$ in the subsequent time slice.

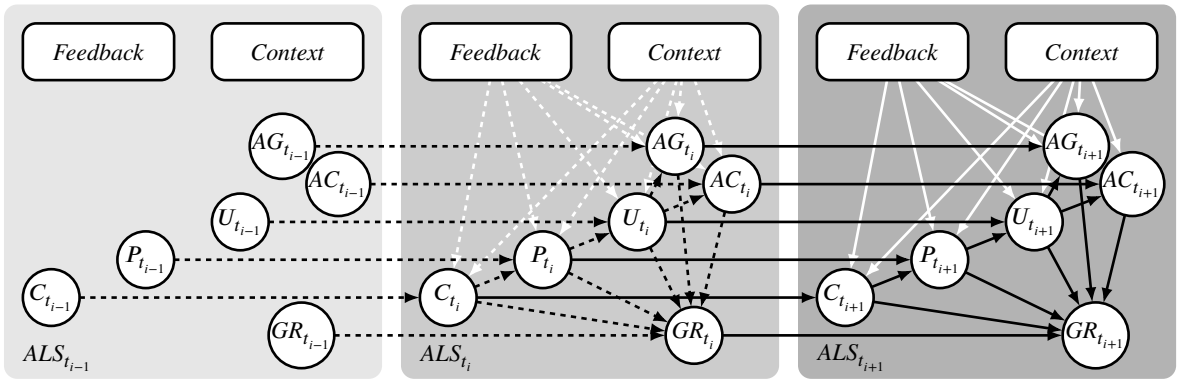


Figure 2: A dynamic two time-slice Bayesian network model unrolling over three steps in time, each corresponding to one dialogue segment. Dashed arrows are disregarded during inference in subsequent time-slices, i.e., variables from time slice $ALS_{t_{i-1}}$ and evidence variable in time slice ALS_{t_i} have no influence on variables in time slice $ALS_{t_{i+1}}$. Posterior distributions of attributed listener state variables in time slice ALS_{t_i} are taken as prior distributions at time t_{i+1} and influence the variables they are connected to in time slice $ALS_{t_{i+1}}$.

Due to the first order Markov assumption, previous time slices ALS_{t_0} to $ALS_{t_{i-1}}$ are not taken into account any more and all connections to them, as well as to all variables X_{t_i} that have no influence into the future, and can be disregarded (dashed lines in Figure 2). The complete history is thus implicitly contained, in accumulated form, in time slice ALS_{t_i}.

In our model, the ALS variables C, P, U, AC, AG, and the groundedness variable GR, are the ones that carry over information between time slices (Figure 2), e.g., understanding at time t_i influences understanding at time t_{i+1} (consequently, variable $U_{t_{i+1}}$ is not only influenced by $P_{t_{i+1}}$, $Feedback_{t_{i+1}}$, and $Context_{t_{i+1}}$, but additionally by U_{t_i}). This is based on the assumption that listener state evolution — and attribution — is usually a gradual process. Indeed, abrupt changes of listener state are often marked by special feedback tokens such as for example *oh* or, in German, *ach* and *ach so*.

Figure 3 simulates the dialogue from example (1) in two contrasting conditions. Once without temporal influences between dialogue segments s_{t_i} and $s_{t_{i+1}}$, based on Buschmeier and Kopp's (2012) static model (Figure 3a); and once with modelled temporal dynamics based on the dynamic model presented above (Figure 3b). Each graph shows how speaker S1's belief state of a specific variable — i.e., the probabilities for each of its elements — changes over time (magenta coloured lines show $P(X = low)$, yellow lines $P(X = medium)$ and cyan coloured lines $P(X = high)$ for $X \in \{P, U, AC, GR\}$). Nine time-steps are shown, each corresponding to one dialogue segment.

In Figure 3a, each feedback event is treated in isolation and independently from the dialogue history. This results in a belief state state that does not change in the beginning, when no feedback is provided by listener U1 (from t_0 to t_2). When U1 provides feedback (from t_3 to t_5 and at t_7), S1's belief state changes abruptly, jumping between rather distant degrees of belief, and returning to the idle state for a brief period of time when no feedback is present (at t_6).

In contrast to this, the dynamic model in Figure 3b, leads to a gradually evolving attributed listener state. In the beginning, when no feedback is provided by U1 (from t_0 to t_2), the belief state shifts towards *low* perception, understanding, acceptance, and groundedness. This changes, cautiously, as soon as feedback is provided at t_3 and grows towards *medium* to *high* with each subsequent feedback signal provided by U1 (at $t_4, t_5,$ and t_7). Notably, at t_6, the belief state does not jump to the initial state, but degrades only slightly while U1 does not provide feedback.

4 Discourse structure and belief state evolution

A question that needs to be addressed is how the attributed listener state in the dynamic model should develop over time, i.e., to what extent and how the belief state ALS_{t_i} influences its successor state $ALS_{t_{i+1}}$. For the example, in Figure 3b, the transitions were assumed to be fixed, that is, the influence $P(X_{t_{i+1}}|X_{t_i})$ of each of the vari-

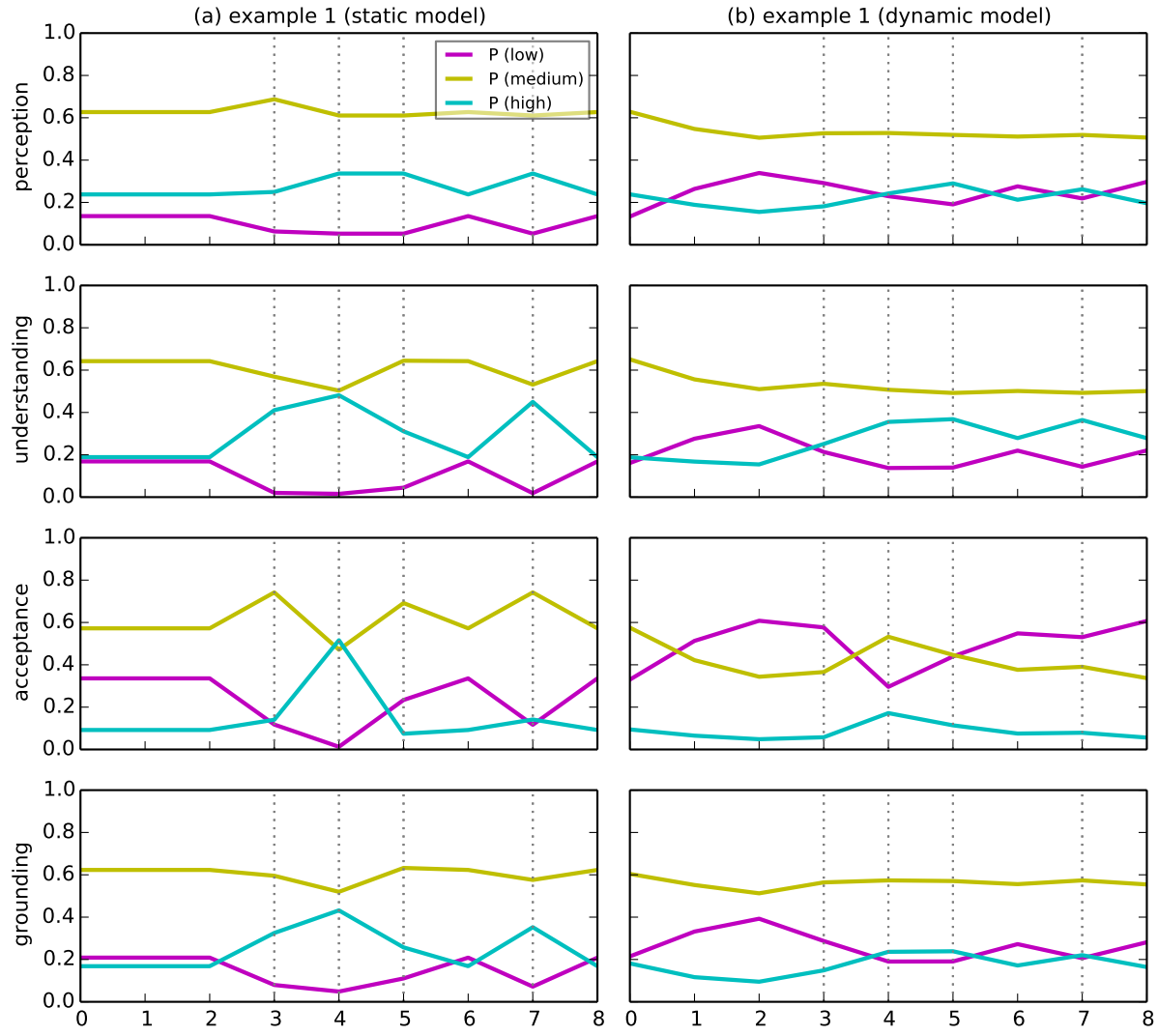


Figure 3: Simulated belief state evolution for example dialogue (1). The graphs show speaker S1's graded belief for the attributed listener state variables P, U, AC, and GR given the feedback provided by listener U1 (dashed vertical lines indicate the exact points in time when feedback occurred). Two conditions are contrasted: (a) without temporal influences between dialogue segments, simulated with Buschmeier and Kopp's (2012) static model; and (b) with temporal influences between dialogue segments, simulated with the two time-slice dynamic Bayesian network model (Figure 2).

ables $X_{t_i} \in \{C_{t_i}, \ldots, GR_{t_i}\}$ on its successor $X_{t_{i+1}} \in \{C_{t_{i+1}}, \ldots, GR_{t_{i+1}}\}$ was fixed for each point in time $t_i \in [t_0, \ldots, t_8]$ (influences among variables varied, i.e., $P(X_{t_{i+1}}|X_{t_i}) \neq P(Y_{t_{i+1}}|Y_{t_i})$ for $X \neq Y$.

This assumption is certainly simplified. As Muller and Prévot (2003) argue, feedback is deeply embedded in the discourse and its relation to the discourse structure is one of its pivotal features. As an example, consider a situation in which at time t_{i+1} either the topic changes, or the narration simply continues. Intuitively, the influence of the speaker's attributed listener state ALS_{t_i} on the attributed listener state $ALS_{t_{i+1}}$ is different in the two situations.

Given a topic change, there is, e.g., little reason to believe that understanding or acceptance as estimated in ALS_{t_i} has much to contribute — i.e., is a good predictor — to understanding and acceptance in $ALS_{t_{i+1}}$ (arguably this also depends on the relatedness of the two topics). In contrast to this, understanding and acceptance as estimated in ALS_{t_i} seems to be very relevant for $ALS_{t_{i+1}}$ in the case where the narration simply continues.

The example indicates that the type of relation between discourse segments — a rhetorical or discourse relation (Asher and Lascarides, 2003) — plays a role in the development of attributed listener state over time. This is in line with the proposal of Stone and Lascarides (2010), who propose a similar influence of discourse relations on grounding, also within an — albeit so far purely theoretical — dynamic Bayesian network model.

As a first approach, we propose that the dynamic model of the listener takes the discourse relation between two consecutive discourse segments into account by simply varying the strength of the influence that a variable X_{t_i} has on a variable $X_{t_{i+1}}$ in the next time-slice. This strength is defined in terms of a weight w that the temporal influence has in relation to the influences of feedback, dialogue context, and other ALS-variables. A weight of $w = 0.5$, for example, results in the influence of X_{t_i} on $X_{t_{i+1}}$ being the the same as the influence that all non-temporal variables have on $X_{t_{i+1}}$. A weight of $0 \leq w < 0.5$ results in temporal influence that is smaller than the influences of the non temporal variables and larger for a weight of $0.5 < w \leq 1$. Concrete weights for individual discourse relations need to be determined empirically.

In practical terms, this approach involves (1) having different dynamic Bayesian network models for each of the discourse relation types, and (2) switching the networks — carrying over the variable assignments and distributions — when proceeding form dialogue segment to dialogue segment.

5 Example applications

In addition to being able to better track the attributed listener state and groundedness, the dynamic minimal model of the listener enables novel applications in artificial conversational agents that were not possible with Buschmeier and Kopp's (2012) static model. Two of these will be sketched in the following.

5.1 Eliciting listener feedback

Listeners do not only produce communicative feedback when they feel the need to inform speakers about their cognitive state of dialogue processing, e.g., if they want to give evidence of understanding or if they do not understand what is said. Often feedback is provided cooperatively in response to 'feedback elicitation cues' of a speaker (Ward and Tsukahara, 2000; Gravano and Hirschberg, 2011). Speakers produce these cues since they have an active interest in how their ongoing utterance is perceived, understood, etc., by their interlocutors, and because it helps them in language production and story telling (Bavelas et al., 2000). This is especially the case in situations where they are uncertain about the listener's cognitive state, even to the extent that they cannot make well-grounded choices in language production. In cases of such an 'information need' (Buschmeier and Kopp, 2014b), elicitation of feedback from the listener is a viable strategy to ensure and achieve an effective dialogue. We propose that the following three criteria — in terms of our model — are indicative of a speaker's information needs (Buschmeier and Kopp, 2014b):

1. The entropy of a variable of interest rises (i.e., the probability distribution across the elements of a variables become more uniform, e.g., when $P(U = low) = 0.33, P(U = medium) = 0.33, P(U = high) = 0.33$) so that the belief state becomes less and less informative.

2. A variable of interest remains static for an extended period of time (e.g., when the listener does not provide feedback).

3. The distance (measured with the Kullback-Leibler divergence) between the probability

distributions of the current state of a variable and a desirable 'reference state' — such as, for example, a state that represents very good understanding — grows beyond a certain acceptable value.

These criteria could in principle be used with the static model of attributed listener state. However, the continuous temporal progression of the belief state makes it possible to identify reliable trends which enable informational needs to be detected early on and with high precision.

5.2 Anticipatory adaptation

A second ability that also builds on the mechanism of identifying trends in the development of the attributed listener state is to adapt language production to anticipate needs of the listener, a mechanism that human speakers use all the time. For this, an artificial agent could simulate the most likely evolution of the dynamic ALS and use this projected next listener state in order to make adaptations in natural language generation that serve as a pre-emptive countermeasure against an expected undesirable cognitive state of the user.

As an example, consider a situation where the agent believes that with every discourse segment the user understood less and less. A simulation that is run for the upcoming segment results in a belief state which shows that this trend is likely to continue. Expecting this state in the dynamic model, now allows the agent to change its original plan — say, to present an additional detail — and instead repeat what has already been said in a different way thus giving the subject matter a different perspective which might help the user understand.

6 Conclusion

In this paper we propose a dynamic Bayesian network-based model for minimal mentalising that tracks the interlocutors' cognitive state with respect to their willingness and ability to perceive, understand, accept, and agree by means of their communicative feedback behaviour. We argued that feedback is a particularly suitable way for listeners to provide evidence of understanding at almost any point in the dialogue, and for speakers to reason about the the listener's cognitive state, as well as to make statements about groundedness. The model can serve as a middle ground between theories that assume representations of full common ground (Clark, 1996) and theories that assume no common ground at all (Pickering and Garrod, 2004).

We extended a previous model of attributed listener state (Buschmeier and Kopp, 2012) with a temporal dimension, showed how the attributed listener state develops while a dialogue unfolds, and illustrated how its progression can be influenced by the structure of the discourse. Finally, we briefly described two relevant and novel applications of the presented model for artificial conversational agents that rely specifically on the model's temporal dynamics and its ability to continuously track the development of the attributed listener state in order to identify trends and project its future development.

Future work will involve an investigation of directionality of the influence of the discourse relations in the dynamic model. A result might be that the flow of information will be reversed given certain discourse relations so that recent evidence of understanding can influences variables in the previous time-slice. We will also implement the mechanisms for feedback cue elicitation and anticipatory adaptation sketched out as applications in an artificial conversational agent and evaluate them in interaction with human users.

A Supplementary material

A data publication containing the model parameters supplements this paper (Buschmeier and Kopp, 2014a). Additionally, the dynamic Bayesian network implementation is publicly available under the GPL 3 license at `http://purl.org/scs/PRIMO`.

Acknowledgements This research is supported by the German Research Foundation (DFG) at the Center of Excellence EXC 277 'Cognitive Interaction Technology' (CITEC).

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26. doi: 10.1093/jos/9.1.1

Nicolas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

David Barber. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, UK.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Per-*

sonality and Social Psychology, 79:941–952. doi: 10.1037/0022-3514.79.6.941

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493. doi:10. 1037/0278-7393.22.6.1482

Sarah Brown-Schmidt. 2012. Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, 27:62–89. doi:10.1080/01690965.2010.543363

Hendrik Buschmeier and Stefan Kopp. 2012. Using a Bayesian model of the listener to unveil the dialogue information state. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 12–20, Paris, France.

Hendrik Buschmeier and Stefan Kopp. 2014a. Dynamic Bayesian model of the listener. Data publication, Bielefeld University, Bielefeld, Germany. doi:10.4119/unibi/2687517

Hendrik Buschmeier and Stefan Kopp. 2014b. When to elicit feedback in dialogue: Towards a model based on the information needs of speakers. In *Proceedings of the 14th International Conference on Intelligent Virtual Agents*, pp. 71–80, Boston, MA, USA.

Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 295–303, Seoul, South Korea.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81. doi:10.1016/j.jml.2003.08.004

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In Aravind K. Joshi et al., (Eds.), *Elements of Discourse Understanding*, pp. 10–63. Cambridge University Press, Cambridge, UK.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK. doi:10.1017/CBO9780511620539

Elizabeth Couper-Kuhlen, Dagmar Barth-Weingarten, et al. 2011. A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 12:1–51.

Alexia Galati and Susan E. Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62:35–51. doi:10.1016/j.jml.2009.09.002

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford University Press, Oxford, UK.

Augustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25:601–634. doi:10.1016/j.csl.2010.10.003

Mattias Heldner, Jens Edlund, and Julia Hirschberg. 2010. Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech 2010*, pp. 3054–3057, Makuhari, Japan.

William S. Horton and Boaz Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59:91–117. doi:10.1016/0010-0277(96)81418-1

Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340. doi:10.1017/S1351324900002539

Philippe Muller and Laurent Prévot. 2003. An empirical study of acknowledgement structures. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, Germany.

Volha Petukhova and Harry Bunt. 2010. Introducing communicative function qualifiers. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pp. 123–131, Hong Kong, China.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226. doi:10.1017/S0140525X04000056

Massimo Poesio and David R. Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13:309–347. doi:10.1111/0824-7935.00042

Christian P. Robert. 1993. Prior feedback: A Bayesian approach to maximum likelihood estimation. *Computational Statistics*, 8:279–294.

Matthew Stone and Alex Lascarides. 2010. Coherence and rationality in grounding. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 51–58, Poznan, Poland.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 38:1177–1207. doi:10.1016/S0378-2166(99)00109-5

Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14:129–182. doi:10.1075/pc.14.1.08war

Victor H. Yngve. 1970. On getting a word in edgewise. In Mary Ann Campbell et al., (Eds.), *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–577. Chicago Linguistic Society, Chicago, IL, USA.

Phrase structure rules as dialogue update rules

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

Abstract

We present a formulation of phrase structure rules in TTR (Type Theory with Records (Cooper, 2012)) as dialogue update rules of a similar kind to those discussed by Ginzburg (2012) and Larsson (2002). This grounds syntax in a theory of events. Apart from unifying syntax with a theory of dialogue processing, there are two main advantages to the proposal: (1) it places constraints on a natural non-abstract theory of syntax which represents linguistic events as they occur and (2) it points us to an account of incremental interpretation in terms of the processing of strings of events in a manner similar to that proposed by Poesio and Traum (1997) and Poesio and Rieser (2010).

1 Introduction

A common view of how dialogue analysis fits into linguistic theory is that dialogue comes as a superordinate structure built on top of syntax, semantics and the other conceptual components of linguistic theory where the kinds of tools used in dialogue analysis seem quite different to what is needed for the other components. I want to suggest that we can turn this around: that everything in linguistic analysis can be thought of in terms of the tools we need for dialogue, that is, tools required for the analysis of communication involving the perception and creation of types of linguistic events and reasoning about updates to information states. And I want to suggest that we can pursue this idea without sacrificing the kind of formal rigour we are able to achieve in more traditional approaches to linguistic analysis.

In this paper we show that we can view phrase structure rules in TTR (Type Theory with Records (Cooper, 2012)) as dialogue update rules. In this way I want to suggest that the foundational notions which run through all the components of linguistic theory have to do with the perception and creation of communicative events in the way that has been discussed in formal theories of dialogue such as Ginzburg (2012) and Larsson (2002). The ideas presented in this paper could be regarded as implicit in their work, although they do not make an explicit connection with phrase structure. There are three aspects of our particular approach which we would like to highlight:

1. Grounding syntax in event perception and creation places intuitive restrictions on what a "natural" syntax is and makes abstract theories of syntax with many inaudible constituents appear as a rather different kind of theory.

2. It also points the way to a view of incremental parsing as information state update in a way which is related to proposals by Poesio and Traum (1997) and Poesio and Rieser (2010).

3. The use of TTR enables us to factor the phrase structure rules into various abstract resources which can be combined. The result gives us a view of universal grammar similar to that of Jackendoff (2002) and Cooper and Ranta (2008) where linguistics universals are regarded as a kind of toolbox from which natural languages select.

The approach we are taking also has a lot in common with that taken by Purver et al. (2010) and Eshghi et al. (2012). There the strategy is to incorporate TTR into Dynamic Syntax. Here the strategy is to incorporate ideas from Dynamic Syntax into TTR. Another related approach which needs to be explored in this connection is represented by Demberg et al. (2013).

We will first present a particular view of update in terms of TTR and then we will show how

phrase structure rules can be considered in these terms. We will not give a detailed introduction to TTR notation[1], although we will use it liberally for the sake of concreteness. However, we will give an intuitive explanation of each formula which should make the ideas accessible to readers unfamiliar with the details of the notation.

We shall consider some of the resources needed to deal with a very simple toy dialogue as in (1).

(1)

User:	Dudamel is a conductor
System:	Aha
User:	Beethoven is a composer
System:	OK

2 Update functions

We will assume that agents do not have complete information about their information state, that is, they reason in terms of *types* of information state (that is, gameboards). The basic intuition behind our reasoning about information state updates can be expressed as in (2).

(2) If $r_i : T_i$, then $r_{i+1} : T_{i+1}(r_i)$

That is, given that we believe that the current information state is of type T_i, then we can conclude that the next information state is of type T_{i+1} which can depend on the current information state. According to this, we can have a hypothesis about the type of the next information state even though we may not know exactly what the current information state is. Thus the dependency in our types provides us with a means for representing underspecification.

This basic rule of inference corresponds to a function from records to record types, a function of type ($T_i \rightarrow$ *RecType*). Such a function is of the form (3).

(3) $\lambda r\!:\!T_i \,.\, T_{i+1}(r)$

Things are a litte more complicated than this, however, because this only represents the change from one information state to another, whereas in fact this change is triggered by an event (speech or otherwise) which bears an appropriate relation to the current information state represented by r. Thus we are actually interested in functions from the current information state to a function from events to the new information state, as in (4).

(4) $\lambda r\!:\!T_i \,.\, \lambda e\!:\!T_e(r) \,.\, T_{i+1}(r, e)$

This is one of a number of ways of characterizing update in this kind of framework. One might for instance think of the type of the speech event as being part of the current information state. Also instead of using an update function one can use a record type with a 'preconditions'-field and an 'effect'-field. Both Ginzburg (2012) and Larsson (2002) have this kind of approach. Our formulation makes explicit that update functions are dependent types, that is functions from objects (including information states and events) to a *type*, in this case for the updated information state. We will see that this makes clear a natural relationship between update functions and phrase structure rules viewed as functions (similar to a categorial grammar approach).

Let us consider the update function which the user could use in order to update her information state after her own utterance of *Dudamel is a conductor*. The function in Figure 1 is modelled on the kind of integration rules discussed in (Larsson, 2002). This function maps information states (records), r, which have a non-empty agenda to a function that maps events to a type of information state. It thus requires that the current information state (the first argument to the function) have a non-empty agenda. The second argument to the function (represented by u) requires the move associated with the speech-event to be of the first type on the agenda in r, the current information state, and also to be an assertion with *SELF* as the speaker. It also requires that the chart associated with this utterance can be interpreted as a move of that type. The requirements on the arguments to the function represent the preconditions. The type that results from applying the function to its arguments represents the effect of the update. This type requires the agenda to be the result of replacing the first type on the agenda in r with an acknowledgement where the speaker is the audience of the assertion move and the audience of the acknowledgement is *SELF*. The content of the acknowledgement is the same as the content of the assertion. That is, what is being acknowledged is the content of the assertion. It furthermore requires

[1]This can be found in Cooper (2012) and in the updated drafts of a manuscript in progress called *Type theory and language: from perception to linguistic communication* to be found on `https://sites.google.com/site/typetheorywithrecords/drafts`.

$$\lambda r{:}\left[\text{ private } : \left[\text{ agenda } : {}_{ne}[\textit{MoveType}(\textit{SELF})]\right]\right]$$
$$\lambda u{:}\left[\begin{array}{lcl}\text{move} & : & \text{fst}(r.\text{private.agenda}) \mathbin{\dot{\wedge}} \left[\text{e:}\begin{bmatrix}\text{sp=}\textit{SELF}\text{:}\textit{Ind}\\ \text{au:}\textit{Ind}\end{bmatrix}\right] \mathbin{\dot{\wedge}} \left[\text{e:}\textit{Assertion}\right]\\ \text{chart} & : & \textit{Chart}\\ \text{e} & : & \text{m-interp(chart,move)}\end{array}\right].$$
$$\left[\text{private:}\left[\text{agenda=}\begin{bmatrix}\text{e:}\textit{Acknowledgement} \mathbin{\dot{\wedge}} \begin{bmatrix}\text{sp=}u.\text{move.e.au:}\textit{Ind}\\ \text{au=}\textit{SELF}\text{:}\textit{Ind}\end{bmatrix}\\ \text{cnt=}u.\text{move.cnt:}\textit{RecType}\\ \text{c}_{\text{cnt}}\text{:content(e,cnt)}\end{bmatrix} \text{:}[\textit{MoveType}(\textit{SELF})] \mid \text{rst}(r.\text{private.agenda})\right]\right.$$
$$\left.\text{shared:}\left[\text{latest-utterance:}\begin{bmatrix}\text{move=}u.\text{move:}\textit{Move}(\textit{SELF})\\ \text{chart=}u.\text{chart:}\textit{Chart}\\ \text{e=}u.\text{e:m-interp(chart,move)}\end{bmatrix}\right]\right]$$

Figure 1

the latest-utterance field to contain the move and chart of the utterance u. The idea is that this function should be used to predict the type of the next information state on the basis of the current information state and the observed event. That is, if we believe the current information state to be of the domain type of the update function and we observe an event of the required type then we reason that the updated information state should be of the type resulting from applying the function to the current information state.

We will now examine how such an update function could be used to reason about an update. Let us suppose that the user considers the current information state to be of type Figure 2.

This represents that the user intends to assert that Dudamel is a conductor represented by the record type $\left[\text{e:conductor(Dudamel)}\right]$. The user also believes that there was no previous utterance and no commitments, i.e. that the planned utterance will be dialogue initial.

Suppose now that the user utters *Dudamel is a conductor* and judges this utterance event u_1 to be an event of type Figure 3.

The user will have more information about the nature of the chart (that is, about what was actually said and how it might be analyzed) than we have represented but we will leave this underspecified.

Clearly in the user's judgement the utterance u_1 fulfils the requirements placed on it by Figure 1 since the move interpretation associated with it is of the type which occurs at the head of the agenda. Note that we are reasoning with this function without actually providing it with an argument since we only have a (hypothesized) type of the current information state, not the actual information state. The crucial judgement is that the type of the current information state is a subtype of the domain type of the function. This is sufficient to allow us to come to a conclusion about the type of the new information state.

According to the update function the next information state must be of the type Figure 4. Note that the speaker in the type on the agenda here is the audience of the original utterance. Thus what is on the agenda is a type of act to be carried out by the interlocutor rather than the *SELF*. This is a way of implementing simple turn-taking in a gameboard approach to dialogue. It also represents the fact that the realization of event types is often a collaborative process. An utterance is not successfully acknowledged if the person who made the original utterance is no longer paying attention, for example.

But we know more about the new information state than what is expressed by the type which results from the update function. Everything we know about the current information state which remains unchanged by the function must be carried over from the current information state. This is related to the frame problem introduced by (McCarthy and Hayes, 1969).[2] We handle this by performing an *asymmetric merge* of the type we have for the current information state with the type resulting from the update function. The asymmetric merge of two types T_1 and T_2 is represented by

[2]For a recent overview of the frame problem see (Shanahan, 2009).

$$\left[\begin{array}{l}\text{private:}\left[\begin{array}{l}\text{agenda=[}\left[\begin{array}{l}\text{e:}\textit{Assertion}\dot{\wedge}\left[\text{sp=}\textit{SELF}\text{:}\textit{Ind}\right]\\ \text{cnt=}\left[\text{e:conductor(dudamel)}\right]\text{:}\textit{RecType}\\ \text{c}_{\text{cnt}}\text{:content(e,cnt)}\end{array}\right]\text{]:[}\textit{RecType}\text{]}\end{array}\right]\\ \text{shared:}\left[\begin{array}{l}\text{latest-utterance:}\textit{Nil}\\ \text{commitments=[]:}\textit{RecType}\end{array}\right]\end{array}\right]$$

Figure 2

$$\left[\begin{array}{lcl}\text{move} & : & \left[\begin{array}{l}\text{e:}\textit{Assertion}\dot{\wedge}\left[\text{sp=}\textit{SELF}\text{:}\textit{Ind}\right]\\ \text{cnt=}\left[\text{e:conductor(Dudamel)}\right]\text{:}\textit{RecType}\\ \text{c}_{\text{cnt}}\text{:content(e,cnt)}\end{array}\right]\\ \text{chart} & : & \textit{Chart}\\ \text{e} & : & \text{m-interp(chart,move)}\end{array}\right]$$

Figure 3

$$\left[\begin{array}{l}\text{private:}\left[\begin{array}{l}\text{agenda=[}\left[\begin{array}{l}\text{e:}\textit{Acknowledgement}\dot{\wedge}\left[\begin{array}{l}\text{sp=}u_1\text{.move.e.au:}\textit{Ind}\\ \text{au=}\textit{SELF}\text{:}\textit{Ind}\end{array}\right]\\ \text{cnt=}u_1\text{.move.cnt:}\textit{RecType}\\ \text{c}_{\text{cnt}}\text{:content(e,cnt)}\end{array}\right]\text{]:[}\textit{RecType}\text{]}\end{array}\right]\\ \text{shared:}\left[\text{latest-utterance:}\left[\begin{array}{l}\text{move=}u_1\text{.move:}\textit{Move}\\ \text{chart=}u_1\text{.chart:}\textit{Chart}\\ \text{e=}u_1\text{.e:m-interp(chart,move)}\end{array}\right]\right]\end{array}\right]$$

Figure 4

$T_1 \boxed{\dot{\wedge}} T_2$. If one or both of T_1 and T_2 are non-record types then $T_1 \boxed{\dot{\wedge}} T_2$ will be T_2. If they are both record types, then for any label ℓ which occurs in both T_1 and T_2, $T_1 \boxed{\dot{\wedge}} T_2$ will contain a field labelled ℓ with the type resulting from the asymmetric merge of the corresponding types in the ℓ-fields of the two types (in order). For labels which do not occur in both types, $T_1 \boxed{\dot{\wedge}} T_2$ will contain the fields from T_1 and T_2 unchanged. In this informal statement we have ignored complications that arise concerning dependent types in record types. Our notion of asymmetric merge is related to the notion of priority unification (Shieber, 1986).

3 Phrase structure rules as update functions

We take signs to be records of the type (5).

$$(5)\quad \left[\begin{array}{lcl}\text{s-event} & : & \textit{SEvent}\\ \text{cnt} & : & \textit{Cnt}\end{array}\right]$$

This represents the pairing of a speech event with content in a Saussurean sign. It does not, however, require the presence of any hierarchical information in the sign corresponding to what in linguistic theory is normally referred to as the *constituent* (or *phrase*) structure of the utterance. To some extent it is arbitrary where we add this information. We could, for example, add it under the label 's-event' ("speech event"). However, it will be more convenient (in terms of keeping paths that we need to refer to often shorter) to add a third field labelled 'syn' ("syntax") at the top level of the sign type as in (6).

$$(6)\quad \left[\begin{array}{lcl}\text{s-event} & : & \textit{SEvent}\\ \text{syn} & : & \textit{Syn}\\ \text{cnt} & : & \textit{Cnt}\end{array}\right]$$

However, as we will see below, *Syn* will require a 'daughters'-field for a string of signs. This means that *Sign* becomes a recursive type. It will be a *basic* type with its witnesses defined by (7).

$$(7)\quad \sigma : \textit{Sign} \text{ iff } \sigma : \left[\begin{array}{lcl}\text{s-event} & : & \textit{SEvent}\\ \text{syn} & : & \textit{Syn}\\ \text{cnt} & : & \textit{Cnt}\end{array}\right]$$

We shall take *Syn* to be the type (8).

(8) $\begin{bmatrix} \text{cat} & : & \mathit{Cat} \\ \text{daughters} & : & \mathit{Sign}^* \end{bmatrix}$

The type *Sign*, as so far defined, can be seen as a *universal resource*. By this we mean that it is a type which is available for all languages. *Cat* is the type of names of syntactic categories. For the purposes of the current toy example we will take the witnesses of *Cat* to be: s ("sentence"), np ("noun phrase"), det ("determiner"), n ("noun"), v ("verb") and vp ("verb phrase"). We will use capitalized versions of these category names to represent types of signs with the appropriate path in a sign type as in (9).

(9) a. $S \equiv \mathit{Sign} \wedge \left[\text{syn:}\left[\text{cat=s:}\mathit{Cat}\right]\right]$

b. $NP \equiv \mathit{Sign} \wedge \left[\text{syn:}\left[\text{cat=np:}\mathit{Cat}\right]\right]$

c. $\mathit{Det} \equiv \mathit{Sign} \wedge \left[\text{syn:}\left[\text{cat=det:}\mathit{Cat}\right]\right]$

d. $N \equiv \mathit{Sign} \wedge \left[\text{syn:}\left[\text{cat=n:}\mathit{Cat}\right]\right]$

e. $V \equiv \mathit{Sign} \wedge \left[\text{syn:}\left[\text{cat=v:}\mathit{Cat}\right]\right]$

f. $VP \equiv \mathit{Sign} \wedge \left[\text{syn:}\left[\text{cat=vp:}\mathit{Cat}\right]\right]$

This means that, for example, (9a) is the type in Figure 5.

We might think that the type *Cat* is a language specific resource and indeed if we were being more precise we might introduce separate types for different languages such as Cat_{eng}, Cat_{swe} and Cat_{tag} for the type of category names of English, Swedish and Tagalog respectively. However, there is a strong intuition that categories in different languages are more or less related. For example, we would not be surprised to find that the categories available for English and Swedish closely overlap (despite the fact that their internal syntactic structure differs) whereas the categories of English and Tagalog have less overlap. (See (Gil, 2000) for discussion.) For this reason we assume that there is a universal resource *Cat* and that each language will have a subtype of *Cat* which specifies which of the categories are used in that particular language. This is related to the kind of view of linguistic universals as a kind of toolbox from which languages can choose which is put forward by Jackendoff (2002) and Cooper and Ranta (2008).

The ontological status of objects of type *Cat* as we have presented them is a little suspicious. Intuitively, categories should be subtypes of *Sign*, as in (9). We have identified signs belonging to these types as containing a particular object in *Cat* in their 'cat'-field. But one might try to characterize such signs in a different way, for example, as fulfilling certain conditions such as having certain kinds of daughters. However, this is not quite enough, for example, for lexical categories, which do not have daughters. We have to have a way of assigning categories to words and we need to create something in the sign-type that will indicate the arbitrary assignment of a category to a word. For want of a better solution we will introduce the category names which belong to the type *Cat* as a kind of "book-keeping" device that will identify a sign-type as being one whose witnesses belong to category bearing that name.

The 'daughters'-field is required to be a string of signs, possibly the empty string, since the type Sign^* uses the Kleene-*, that is the type of strings of signs including the empty string, ε. Lexical items, that is words and phrases which are entered in the lexicon, will be related to signs which have the empty string of daughters. We will use *NoDaughters* to represent the type $\left[\text{syn:}\left[\text{daughters=}\varepsilon\text{:}\mathit{Sign}^*\right]\right]$.

If T_{phon} is a type (normally a phonological type, that is, $T_{\text{phon}} \sqsubseteq \mathit{Phon}$) and T_{sign} is a type (normally a sign type, that is,$T_{\text{sign}} \sqsubseteq \mathit{Sign}$, then we shall use Lex(T_{phon}, T_{sign}) to represent Figure 6. This means, for example, that Figure 7(a) represents the type in Figure 7(b) which, after spelling out the abbreviations, can be seen to be the type in Figure 7(c). We can think of 'Lex' as the function in (10)[3]

(10) $\lambda T_1\text{:}\mathit{Type}$
$\lambda T_2\text{:}\mathit{Type}$.
$T_1 \wedge \left[\text{s-event:}\left[\text{e:}T_2\right]\right] \wedge \mathit{NoDaughters}$

This function, which creates sign types for lexical items in a language, associating types with a syntactic category, can be seen as a universal resource. We can think of it as representing a (somewhat uninteresting, but nevertheless true) linguistic universal: "There can be speech events of given types which have no daughters (lexical items)".

[3]We are using the notational convention for function application as used, for example, by (Montague, 1973) that if f is a function $f(a, b)$ is $f(b)(a)$.

$$\begin{bmatrix} \text{s-event} & : & \begin{bmatrix} \text{e-loc} & : & \textit{Loc} \\ \text{sp} & : & \textit{Ind} \\ \text{au} & : & \textit{Ind} \\ \text{e} & : & \textit{Phon} \\ \text{c}_{\text{loc}} & : & \text{loc(e,e-loc)} \\ \text{c}_{\text{sp}} & : & \text{speaker(e,sp)} \\ \text{c}_{\text{au}} & : & \text{audience(e,au)} \end{bmatrix} \\ \text{syn} & : & \begin{bmatrix} \text{cat=s} & : & \textit{Cat} \\ \text{daughters} & : & \textit{Sign}^* \end{bmatrix} \\ \text{cnt} & : & \textit{Cnt} \end{bmatrix}$$

Figure 5

$$T_{\text{sign}} \mathbin{\dot{\wedge}} \left[\text{s-event:}\left[\text{e:}T_{\text{phon}}\right]\right] \mathbin{\dot{\wedge}} \textit{NoDaughters}$$

Figure 6

a. Lex("Dudamel", *NP*)

b. $\textit{NP} \mathbin{\dot{\wedge}} \left[\text{s-event:}\left[\text{e:"Dudamel"}\right]\right] \mathbin{\dot{\wedge}} \textit{NoDaughters}$

c.
$$\begin{bmatrix} \text{s-event} & : & \begin{bmatrix} \text{e-loc} & : & \textit{Loc} \\ \text{sp} & : & \textit{Ind} \\ \text{au} & : & \textit{Ind} \\ \text{e} & : & \text{"Dudamel"} \\ \text{c}_{\text{loc}} & : & \text{loc(e,e-loc)} \\ \text{c}_{\text{sp}} & : & \text{speaker(e,sp)} \\ \text{c}_{\text{au}} & : & \text{audience(e,au)} \end{bmatrix} \\ \text{syn} & : & \begin{bmatrix} \text{cat=np} & : & \textit{Cat} \\ \text{daughters=}\varepsilon & : & \textit{Sign}^* \end{bmatrix} \\ \text{cnt} & : & \textit{Cnt} \end{bmatrix}$$

Figure 7

The lexical resources needed to cover our example fragment is given in (11).

(11) Lex("Dudamel", *NP*)
Lex("Beethoven", *NP*)
Lex("a", *Det*)
Lex("composer", *N*)
Lex("conductor", *N*)
Lex("is", *V*)
Lex("ok", *S*)
Lex("aha", *S*)

The types in (11) belong to the specific resources required for English. This is not to say that these resources cannot be shared with other languages. Proper names like *Dudamel* and *Beethoven* have a special status in that they can be reused in any language, though often in modified form, at least in terms of the phonological type with which they are associated without this being perceived as quotation, code-switching or simply showing off that you know another language.

Resources like (11) can be exploited by update rules. If Lex(T_w, C) is one of the lexical resources available to an agent A and A judges an event e to be of type T_w, then A is licensed to update their gameboard with the type Lex(T_w, C). Intuitively, this means that if the agent hears an utterance of the word "composer", then they can conclude that they have heard a sign which has the category noun. This is the beginning of *parsing*. The licensing condition corresponding to lexical resources like (11) is given in Figure 8. We will

return below to how this relates to gameboard update. Figure 8 says that an agent with lexical resource Lex(T, C) who judges a speech event, u, to be of type T is licensed to judge that there is a sign of type Lex(T, C) whose 's-event.e'-field contains u.

Strings of utterances of words can be classified as utterances of phrases. That is, speech events are hierarchically organized into types of speech events. Agents have resources which allow them to reclassify a string of signs of certain types ("the daughters") into a single sign of another type ("the mother"). So for example a string of type *Det*⌒*N* (that is, a concatenation of an event of type *Det* and an event of type *N*) can lead us to the conclusion that we have observed a sign of type *NP* whose daughters are of the type *Det*⌒*N*. The resource that allows us to do this is a rule which we will model as the function in (12a) which we will represent as (12b).

(12) a. $\lambda u : \textit{Det}^\frown\textit{N}\ .$
$\textit{NP} \mathbin{\dot\wedge} \left[\text{syn:}\left[\text{daughters=}u\text{:}\textit{Det}^\frown\textit{N}\right]\right]$

b. RuleDaughters(*NP*, *Det*⌒*N*)

'RuleDaughters' is to be the function in Figure 9. Thus 'RuleDaughters', if provided with a subtype of $\textit{Sign}^+$ and a subtype of *Sign* as arguments, will return a function which maps a string of signs of the first type to the second type with the restriction that the daughters field is filled by the string of signs. 'RuleDaughters' is one of a number of sign type construction operations which we will introduce as universal resources which have the property of returning what we will call a sign combination function. The licencing conditions associated with sign combination functions are as characterized in Figure 10. This means, for example, that if you categorize a string of signs, u, as being of type *Det*⌒*N* then you can conclude that there is a sign of type *NP* with the additional restriction that its daughters are u.

'RuleDaughters' takes care of the 'daughters'-field but it says nothing about the 's-event.e'-field, that is the phonological type associated with the new sign. This should be required to be the concatenation of all the 's-event.e'-fields in the daughters. If $u : T^+$ where T is a record type containing the path π, we will use $\text{concat}_i(u[i].\pi)$, the concatenation of all the values $u[i].\pi$ for each element in the string u in the order in which they occur in the string. We can now formulate the function ConcatPhon as in Figure 11. ConcatPhon will map any string of speech events to the type of a single speech event whose phonology (that is the value of 's-event.e') is the concatenation of the phonologies of the individual speech events in the string.

We want to combine the function in Figure 11 with a function like that in (12). We do this by merging the domain types of the two functions and also merging the types that they return. This is shown in Figure 12(a) which in deference to standard linguistic notation for phrase structure rules could be represented as Figure 12(b).[4] In general we say that if $C, C_1, \ldots, C_n$ are category sign types as in (9) then $C \longrightarrow C_1 \ldots C_n$ represents RuleDaughters(C, $C_1{}^\frown \ldots {}^\frown C_n$) $\mathbin{\dot\wedge}$ ConcatPhon where for any type returning functions $\lambda r : T_1\ .\ T_2(r)$ and $\lambda r : T_3\ .\ T_4(r)$ $\lambda r : T_1\ .\ T_2(r) \mathbin{\dot\wedge} \lambda r : T_3\ .\ T_4(r)$ denotes the function $\lambda r : \dot T_1 \mathbin{\dot\wedge} T_3\ .\ T_2(r) \mathbin{\dot\wedge} T_4(r)$. Thus the function in Figure 12 can be represented in a third way as in Figure 13. The hope is that the ability to factorize rules into "bite-size" components will enable us to build a theory of resources that will allow us to study them in isolation and will also facilitate the development of theories of learning. It gives us a clue to how agents can build new rules by combining existing components in novel ways. It has implications for universality as well. For example, while the rule *NP* $\longrightarrow$ *Det N* is not universal (though it may be shared by a large number of languages), ConcatPhon is a universally available rule component, albeit a trivial universal, which says that you can have concatenations of speech events to make a larger speech event.

The rules associated with our small grammar are given by (13).

(13) *S* $\longrightarrow$ *NP VP*
NP $\longrightarrow$ *Det N*
VP $\longrightarrow$ *V NP*

4 Conclusions

It may seem that we have done an awful lot of work to arrive at simple phrase structure rules. Some readers might wonder why it is worth all this trouble to ground the rules in a theory of events

[4] Note that '$\longrightarrow$' used in the phrase structure rule in Figure 12(b) is not the same arrow as '$\rightarrow$' which is used in our notation for function types. We trust that the different contexts in which they occur will help to distinguish them.

If Lex(T, C) is a resource available to agent A, then for any u, $u :_A T$ licenses $:_A$ Lex(T, C) $\wedge_{:} \left[\text{s-event:} \left[\text{e=}u\text{:}T_1\right]\right]$

Figure 8

$$\lambda T_1 : \textit{Type}$$
$$\quad \lambda T_2 : \textit{Type} \; .$$
$$\quad\quad \lambda u : T_1 \; . \; T_2 \wedge_{:} \left[\text{syn:} \left[\text{daughters=}u\text{:}T_1\right]\right]$$

Figure 9

If $f : (T_1 \rightarrow Type)$ is a sign combination function available to agent A, then for any u, $u :_A T_1$ licenses $:_A f(u)$

Figure 10

$$\lambda u\text{:} \left[\text{s-event:} \left[\text{e:}\textit{Phon}\right]\right]^+ .$$
$$\left[\text{ s-event } : \left[\text{ e=concat}_i(u[i].\text{s-event.e}) : \textit{Phon} \right] \right]$$

Figure 11

and action when what we come up with in the end is something that can be expressed in a standard notation which is one of the first things that a student of syntax learns. One reason has to do with our desire to explore the relationship between the perception and processing of non-linguistic events and speech events. Another reason has to do with placing natural constraints on syntax. By grounding syntactic structure in types of events we provide a motivation for the kind of discussion in (Cooper, 1982). An abstract syntax which proposes constituent structure which does not correspond to speech events is not grounded in the same way and thus presents a different kind of theory. The abstraction lies in the nature of the types used to classify strings, rather than abstract elements in the strings themselves. A third reason is that it points to a way of thinking of parsing in TTR as incremental updating of an information state similar to the kind of proposals that have been made in PTT (Poesio and Traum (1997) and Poesio and Rieser (2010)). We have not integrated our view of syntax with compositional semantics and dialogue update rules here. This is, however, done in the work in progress cited in footnote 1.

Acknowledgments

This paper was supported in part by VR project 2009-1569, Semantic analysis of interaction and coordination in dialogue (SAICD). I am grateful to three anonymous referees for insightful comments.

a. $\lambda u : \textit{Det}^\frown N \wedge_{:} \left[\text{s-event:} \left[\text{e:}\textit{Phon}\right]\right]^+ .$
$\quad \textit{NP} \wedge_{:} \left[\text{syn:} \left[\text{daughters=}u\text{:}\textit{Det}^\frown N\right]\right]$
$\quad\quad \wedge_{:} \left[\text{s-event:} \left[\text{e=concat}_i(u[i].\text{s-event.e)}\text{:}\textit{Phon}\right]\right]$

b. $\textit{NP} \longrightarrow \textit{Det N}$

Figure 12

$$\text{RuleDaughters}(\textit{NP}, \textit{Det}^\frown N) \wedge_{:} \text{ConcatPhon}$$

Figure 13

References

Robin Cooper and Aarne Ranta. 2008. Natural Languages as Collections of Resources. In Robin Cooper and Ruth Kempson, editors, *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*, volume 1 of *Communication, Mind and Language*, pages 109–120. College Publications, London.

Robin Cooper. 1982. Binding in wholewheat* syntax (*unenriched with inaudibilia). In Pauline Jacobson and Geoffrey K. Pullum, editors, *The Nature of Syntactic Representation*, volume 15 of *Synthese Language Library*. Reidel Publishing Company.

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.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational Interactions: Capturing Dialogue Dynamics. In Staffan Larsson and Lars Borin, editors, *From Quantfication to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications.

David Gil. 2000. Syntactic categories, cross-linguistic variation and universal grammar. In Petra M. Vogel and Bernard Comrie, editors, *Approaches to the typology of word classes*, volume 23 of *Empirical approaches to language typology*. Mouton de Gruyter, Berlin.

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

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

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

J. McCarthy and P. J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502.

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.

M. Poesio and H. Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1(1):1–89.

M. Poesio and D. Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. 2010. Splitting the *I*s and Crossing the *You*s: Context, Speech Acts and Grammar. In Paweł Łupkowski and Matthew Purver, editors, *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–50, Poznań. Polish Society for Cognitive Science.

Murray Shanahan. 2009. The frame problem. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/win2009/entries/frame-problem/`, winter 2009 edition.

Stuart Shieber. 1986. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Publications, Stanford.

Numerical expressions, implicatures and imagined prior context

Chris Cummins
Department of Linguistics and English Language
University of Edinburgh
c.r.cummins@gmail.com

Abstract

Pragmatic interpretations are, by definition, influenced by contextual factors. Research in experimental semantics and pragmatics has suggested that participants, when presented with fragments of discourse, draw inferences about the nature of the prior context and use these inferences to shape their interpretation of the target materials. This has both methodological and theoretical implications. Focusing on the domain of numerical expressions, I present an experiment that aims to elucidate the effect of participants imagining a particular prior context (specifically, one in which a given numeral is salient). I show that this expectation influences pragmatic interpretation in a classically predictable way. I further argue that the effect of 'imagined prior context' might be responsible for a sizeable portion of the unexpected variability exhibited between participants in typical pragmatic experiments.

1 Introduction

A substantial body of research in experimental semantics and pragmatics has addressed the generation of so-called scalar implicatures (SIs). SIs constitute a special case of the more general quantity implicature, in which – following the analysis of Grice (1989) – hearers use the speaker's utterance to draw inferences about the falsity of logically stronger alternatives that could have been uttered instead. SIs specifically rely on the existence of informational scales, comprising terms which belong to the same semantic field but differ in informational strength.

The canonical example of scalar implicature, both historically and in the current experimental literature, involves the scale <some, all>. Taking "some" to possess purely existential semantic meaning, "all" entails "some", and in that sense is informationally stronger (across a wide range of possible contexts of use). Consequently, the hearer of (1) is argued to be able to recover the implicature (2), as first observed by Mill (1865).

(1) I saw some of your children today.

(2) The speaker saw some but not all of the addressee's children today.

The availability of such an implicature relies upon a number of auxiliary assumptions, including that the speaker is knowledgeable about the stronger proposition (as already pointed out by Mill) and potentially that the stronger proposition is relevant to the discourse purpose (see for example Breheny et al. 2006). However, those assumptions being met, implicatures should be recoverable by any competent user of language. Indeed, on a Gricean analysis, they are an aspect of intentional communication: the speaker of (1) explicitly intends to convey the meaning "some but not all", and the work of the hearer is merely to recover this intention. In that sense, the ability to recover implicatures is a necessary part of a language user's communicative competence (at least if we accept the general characterization of linguistic communication as 'intentional').

From this point of view, it is unsurprising that developmental research has documented that young children appear to lack facility with implicatures (Papafragou & Musolino 2003, Guasti et al. 2005, and many others). However, it is profoundly surprising that numerous adult studies have documented acceptance rates for the implicature "some" +> "not all" that are far from maximal (Noveck 2001, Bott & Noveck 2004, Guasti et al. 2005, etc.)

This cannot readily be attributed to deficiencies in the specific scale being tested, <some, all>. Of course, this scale may indeed be deficient in some respect, but comparative research suggests that it is nevertheless among the strongest and most reliable of the posited implicatural scales (van Tiel et al. in prep.) Hence, if the <some, all> scale lacks explanatory value, we might argue that the same is true of scalar implicature in general.

A less radical alternative account for the variability in performance, both between and within tasks, is that it is driven by contextual factors. Depending on the precise nature of the task, an underinformative choice of expression – such as

saying "some" when in fact "all" is the case – might be acceptable to a greater or lesser degree. For instance, we might expect that underinformative "some" would be less acceptable if the task is understood to involve giving the best possible description, but more acceptable if the task merely involves making any true statement. The nature of the judgment that participants are obliged to make could also exert an influence here, as for instance in Katsos and Bishop's (2011) study. They demonstrate that children aged 5 reliably accept (and adults reliably reject) descriptions with "some" given to situations with "all", when the response condition is effectively binary (yes/no). However, when the response condition is ternary (in effect, good, bad or medium), children and adults alike reliably assign the intermediate rating to underinformative descriptions with "some". This suggests, as Katsos and Bishop argue, that the children's behaviour in the binary condition does not reflect their lack of awareness of the shortcomings of the tested utterances. Rather, it seems to reflect an unwillingness on the children's part to reject utterances on this basis, an unwillingness that adults do not share.

Can we invoke a contextual explanation to deal with within-task variability, though? In such cases, the presented context is the same for all participants, yet the observed behaviour varies. The only possible contextual explanation for this is that participants – in addition to taking into account the provided context – are imagining more elaborate and detailed prior contexts for the utterances, and that these contexts differ between participants, for instance in the level of accuracy or informativeness that they require the following utterance to exhibit.

The idea that participants in experiments of this kind might conjure up richer contexts for interpretation is not a new one – Breheny et al. (2006), for instance, explicitly note this possibility. However, it appears that relatively little attention has been paid to documenting directly whether this phenomenon exists, and if so, whether or not it is widespread. This omission is surprising given the potential methodological importance of such work for experimental semantics and pragmatics. As a research area, experimental pragmatics grapples directly with this issue, in that the object of study is the meanings of real-life utterances produced in particular contexts, but the experimental research that addresses this question relies heavily on artificially constructed materials which are necessarily often presented in relatively impoverished contexts. In experiments, it is more typical to present a single conversational turn or a question-answer pair than a full dialogue, and it is hard to exclude the possibility that participants may make assumptions about the higher-order discourse purpose or the content of previous turns to which they were not privy.[1]

Indeed, even our theoretical intuitions about pragmatic meanings may be informed by speculation about the likely context of utterance, even when this is not treated in a systematic fashion by theory. Even the uncontroversial intuition that "some" can convey "not all" relies on the assumption that the stronger proposition "all" might have been relevant, given the prior discourse context, in circumstances in which "some" can be uttered, an assumption that in turn relies on a notion of relevance that is somewhat elusive. For less frequently occurring forms, such as those discussed in the following section, the problem may be more severe, as the form may effectively carry more information about its own likely context of utterance than is generally acknowledged.

In this paper, I make a preliminary attempt at addressing the issue of 'imagined prior context' experimentally. In doing so, I focus on pragmatic enrichments within the numerical domain, a decision that I attempt to motivate in the following section.

2 Implicatures from numerical expressions

The domain of numerical expressions appears to be a fertile one for pragmatic enrichment. A popular analysis of numeral meaning holds that numbers are lower-bounded on their semantics and acquire exact meanings pragmatically through implicature (although see Breheny 2008 for a critical discussion of this proposal). More recently, Cummins, Sauerland and Solt (2012) demonstrate the availability of pragmatic enrichments, apparently due to quantity implicature, from expressions of the form "more than *n*".

[1] An anonymous reviewer raised the general and very important question of what artificial experiments of this kind can tell us about natural communication. I have no space here to offer a manifesto for experimental pragmatics, as practised at the sentence level. However, I would argue that both the process of enriching weak scalar meanings and the process of inferring non-shared prior context are highly likely to be relevant to natural communication. Nevertheless, my immediate concern here is just to try to disentangle those two processes in laboratory tasks.

They also argue that these enrichments are conditioned by numeral salience.

To take a specific example, Cummins et al. (2012) show experimentally that quantifying sentences such as (3) are considered to convey additional meanings to the effect that, for instance, (4) or (5).

(3) I have more than 60 CDs.
(4) I do not have more than 80 CDs.
(5) I do not have more than 100 CDs.

The available implicatures are argued to depend upon the salience of the numeral concerned. That is, Cummins et al.'s account explains the absence of an implicature to the effect that (6) is false, given the utterance (3), by arguing that (6) is independently disfavoured on the basis of using a non-salient number. Hence, the speaker's decision to utter (3) rather than (6) can be explained just as a preference for using the number 60 rather than 61, and consequently there is no need for the hearer to postulate that the speaker is unable to commit to the truth of the assertion (6). For this reason, the implicature not-(6) is predicted to be unavailable, as is borne out experimentally.

(6) I have more than 61 CDs.

Whether or not this particular account is along the right lines, Cummins et al.'s data seems strongly to suggest that implicatures are available in principle from utterances containing "more than *n*" for numeral *n*. Moreover, for certain values of *n*, a wide range of different implicatures appear to be available, depending on the preferences of the individual participant. A given instance of "more than 100" can be construed as conveying "not more than 110", "not more than 125", "not more than 150" or "not more than 200". Hence, just like the some/all case, there is considerable variation between participants as to whether specific pragmatic enrichments are endorsed. Indeed, the picture is more colourful in the numerical case, inasmuch as a greater number of distinct candidate implicatures (or sets of implicatures) are endorsed by different participants, but again the reasons for this are not clearly understood. Moreover, as noted by Fox and Hackl (2006), such implicatures are not observed in the cases of small cardinal quantities ("more than two people" does not implicate "not more than three people"), which is another fact requiring explanation.

For numerical expressions, as opposed to other expressions of quantity, it also seems more feasible to be able to ask participants direct questions about the choice of expression. Given an utterance such as (3), the question "Do you think that the specific number 60 was important for some reason?" seems perfectly reasonable and is not a leading question. By contrast, given an utterance such as (1), the question "Do you think that the specific quantity 'some' was important for some reason?" seems less natural.

For all these reasons, I would argue that the domain of numerical expressions is a particularly convenient testbed for the hypothesis sketched out in the introduction: namely that the variability between participants in their generation of implicatures is partly explicable in terms of the different prior contexts that they imagine. The experiment in the following section sets out to investigate this claim.

3 Experiment: implicatures and inferences about prior context

In this experiment, participants read sentences containing numerically-quantified expressions, and were asked a set of questions about each sentence. The aim was to examine simultaneously whether the kind of implicature predicted by Cummins et al. (2012) was available, whether the reader inferred that the specific number was being used for a particular reason, and whether (as predicted by, for instance, a traditional Gricean pragmatic account) these two forms of inference were inversely correlated in strength.

3.1 Materials

12 sentences containing numerically-quantified expressions were sampled from the BNC (BNC, 2007). These comprised one instance each of "more than 60", "more than 70", "more than 80", "more than 90", "at least 60", "at least 70", "at least 80", "at least 90", "more than one", "more than two", "more than three", and "more than four". The usage of each expression was cardinal and related to the number in question: instances such as "more than 50 per cent", "more than 60 million", and "more than 70 metres" were excluded from consideration. Bearing in mind Cummins et al.'s (2012) findings about the presence of prior context, sentences were also excluded from consideration if the preceding sentence contained a numeral (or if there was no preceding sentence, i.e. the sentence in question was the beginning of a text). However, the preceding sentences were in any case not presented to participants in this study.

Instances of "more than/at least *n*" for non-round *n* are rare in the BNC and no appropriate examples of cardinal usage, respecting the above criteria, could be located. For this reason, non-round conditions were created by replacing the above numbers with non-round numbers of the same order of magnitude: 60 with 58, 70 with 77, 80 with 86, and 90 with 93.[2]

Two lists were created, each comprising 12 items in pseudorandom order. The four small-number "more than" sentences were presented on both lists. For the remaining items, the design balanced between round (original) and non-round (replacement) numbers. Thus, version 1 contained sentences with "more than 60", "more than 77", "more than 86" and "more than 90", whereas version 2 contained those same sentences with "more than 58", "more than 70", "more than 80" and "more than 93". For "at least", the reverse was true: version 1 contained "at least 58/70/80/93" and version 2 contained "at least 60/77/86/90". In this way, each participant saw each sentence and each number only once. The sentences used are shown in Appendix A.

For each item, participants were asked to judge four statements on a five-point Likert scale rated from "very unlikely" (1) to "very likely" (5). The first statement concerned the availability of a specific implicature predicted by Cummins et al. (2012); for instance, where the text identified the existence of "more than 70 volumes", statement (i) was "In the speaker's opinion, the actual number of volumes is less than 80". Statement (ii) was "The speaker said [more than 70] because that was the most informative statement possible". Statement (iii) was "The speaker said [more than 70] because that was a convenient approximation". Statement (iv) was "The speaker said [more than 70] because the specific number [70] was important for some reason".

3.2 Participants

Participants were recruited via Amazon Mechanical Turk. The conditions were fielded on separate days in April 2014. 17 participants completed version 1 of the experiment and 14 participants completed version 2.

[2] An anonymous reviewer observes that the construction of materials in this way could be seen as an advantage, in that it reduces the amount of irrelevant variance. However, for the present purposes, I consider this a potential disadvantage, as I must then assume without proof that the resulting materials are in fact pragmatically felicitous.

3.3 Results

As no major differences were observed between the results from the two conditions, they are pooled and considered together in what follows. Table 1 presents the mean ratings (and SDs) for each of the test conditions.

	(i)	(ii)	(iii)	(iv)
More than				
Round	3.46 (1.30)	3.44 (1.15)	4.08 (1.06)	2.98 (1.09)
Non-round	3.63 (1.12)	3.68 (1.04)	3.29 (1.23)	3.11 (1.27)
Small	2.02 (1.27)	3.43 (1.13)	3.29 (1.20)	3.58 (1.24)
At least				
Round	3.37 (1.41)	3.67 (1.04)	3.90 (0.94)	3.10 (1.16)
Non-round	3.27 (1.38)	3.87 (1.09)	3.21 (1.33)	3.27 (1.26)

Table 1: Mean ratings (and SDs) for each quantifier and number condition

Considering the mean responses for each tested item within each category (i.e. the means by-sentence), the ratings for (i) and (iv) are strongly negatively correlated (Pearson's $r = -0.67$). These mean ratings are tabulated in full in Appendix B. Planned comparisons via t-tests indicate that the ratings in the "more than" condition with respect to statement (i) are lower for small numbers than for either round or non-round numbers, and with respect to statement (iv) are higher for small numbers than for either round or non-round numbers (all $p < 0.01$).

3.4 Discussion

The existence of a strong negative correlation between judgments of statements (i) and (iv) seems to suggest that, where participants infer that specific numerals are being used for a particular reason, they are disinclined to infer the otherwise-predicted pragmatic enrichment. This appears to concur with the predictions of Cummins et al. (2012). Recall that the availability of an enrichment of the kind canvassed in (i) requires that a stronger alternative assertion was available to the speaker, and that this alternative was not selected purely on the grounds of its falsity. By contrast, where a specific numeral is

chosen because it is somehow intrinsically special (as evidenced by a high rating for statement (iv)), the informationally weaker assertion may be preferable to informationally stronger alternatives, on the basis that these stronger alternatives would fail to use the "special" number. Consequently, the speaker's decision to use the informationally weaker assertion should not convey anything about the truth-value of the informationally stronger alternative in this particular case.

Delving into the specific conditions, the results suggest that participants are strongly disinclined to endorse the candidate implicatures arising from the small number conditions "more than two/three/four/five" (respectively, "not more than three/four/five/six". This is unsurprising – these implicatures have been widely assumed to be unavailable (see for example Fox and Hackl 2006), at least in cardinal contexts. More strikingly, these expressions give rise to clear judgments that the numbers in question are likely to be contextually salient (as shown by their high ratings on statement (iv)), even in the absence of any explicit contextual support for this claim.

The unavailability of these implicatures could be attributed to several distinct causes. One possibility (explored by Fox and Hackl 2006) is that expressions of the form "more than *n*" systematically fail to give rise to implicatures: however, this appears to over-predict, in the light of Cummins et al.'s data. Another possibility is that the implicatures are blocked as a consequence of their communicative oddness: if "more than two" implicated "not more than three", these premises would together entail "exactly three", which could be much more easily communicated in other words. This would also account for the intuition that "more than two" gives rise to implicatures in measurement contexts, with "more than two metres" implicating "not more than three metres". However, the results of this experiment could be taken to support a third explanation, namely that the systematic lack of implicatures from expressions such as "more than two" stems from the fact that these expressions trigger strong expectations that the specific numeral used was used for a particular reason. A rational hearer who held such an expectation should be unwilling to draw quantity implicatures. For instance, suppose that the hearer assumes "more than two" is being used because "two" is an especially salient number. It follows that the more informative "more than three" might not be a better alternative, even if it is true, on the basis that it fails to use this salient number "two". The hearer should conclude that the use of "more than two" rather than "more than three" does not necessarily signal the speaker's unwillingness to commit to the truth of that latter, stronger proposition.

Of course, this explanation is only tenable if sentences involving "more than two" in cardinal contexts are restricted in their distribution. They would be predicted to be admissible in situations in which the number "two" is salient, or can be presumed to be salient: in such situations, the implicature "not more than three" would be blocked for the reason discussed above. "More than two" would also be predicted to be admissible in situations in which the speaker is not knowledgeable about the truth of stronger propositions, in which case the implicature would fail to arise for standard reasons (this epistemic assumption being essential for implicature on the traditional account). However, "more than two" would be predicted not to be admissible in situations in which the speaker is knowledgeable about the precise value and in which the number "two" is not especially salient. Examples discussed in the literature such as (7), in which the speaker turns out to be knowledgeable about the precise value, appear strongly to invite the inference that having "two children" constitutes a threshold of some kind (e.g. for entitlement for benefits). However, the question remains open as to whether all examples of "more than two" in cardinal quantificational contexts actually have this property.

(7) John has more than two children; in fact, he has five.

In the case of large round numbers, participants are inclined to draw the pragmatic enrichment, endorsing statement (i). This replicates the findings of Cummins et al. (2012). Moreover, participants strongly endorsed statement (iii) in this case (the rating exceeding that for both other conditions; t-tests, $p < 0.01$). This suggests that these utterances are regarded as convenient approximations rather than attempts to use specific numbers; hence, implicatures should be available. This expectation seems to be borne out.

Large non-round numbers behave similarly to large round numbers in this experiment, but were numerically rated higher with respect to both statement (iv) and statement (i). They scored somewhat lower on (iii), perhaps indicating that they are not as 'convenient' an approximation as round numbers; and slightly higher on (ii), suggesting that they can be perceived as optimally

informative. This fits with the assumption that the use of non-round numbers permits greater precision but is associated with additional cognitive costs. It is tempting to hypothesize that the large non-round numbers constitute an intermediate case between round and small numbers in this experiment, and that the speaker who uses such a number is presumed both to be deliberately using a specific number and to be attempting to convey an implicature. This would be conceivable if the hearer presumes that the speaker might prefer to use some specific number, but may not be willing to sacrifice a great deal of informativeness in order to do so: for example, even if 83 is a salient number, a speaker might use "more than 100" in preference to "more than 83" if they know the informationally stronger statement to be true. However, more work is required both in order to determine whether speakers actually exhibit this kind of preference, and – independently of that – whether hearers perceive that speakers are going to exhibit this kind of preference, and can modulate their interpretations of quantity expressions accordingly.

4 Conclusion

The experiment presented in this paper represents a preliminary attempt to explore the idea that numerically-quantified expressions might signal information about the prior context against which they should be interpreted, even when this prior context is not provided. The results of the experiment do appear to suggest that this is the case: participants spontaneously infer that specific numbers (of particular kinds) are contextually salient, purely on the basis of their usage. The implicatures recovered by participants appear to be modulated by this perception of contextual salience, although it is not possible to infer the existence of a causal relationship on the basis of this experiment.

Based on these findings, it is tempting to posit that at least some of the variability between participants, documented in experiments on quantity implicature, might be attributed to differences in the way in which they infer details of the context of utterance. The domain of number represents a convenient testbed for this approach, but in principle the hypothesis makes predictions about a much wider range of situations. Future work will aim both to broaden and deepen the experimental exploration of this area.

References

BNC. 2007. *The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Lewis Bott and Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3): 437-457.

Richard Breheny, Napoleon Katsos and John N. Williams (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3): 434-463.

Richard Breheny. 2008. A new look at the semantics and pragmatics of Numerically Quantified Noun Phrases. *Journal of Semantics*, 25(2): 93-139.

Chris Cummins, Uli Sauerland and Stephanie Solt. 2012. Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy*, 35: 135-169.

Danny Fox and Martin Hackl. 2006. The universal density of measurement. *Linguistics and Philosophy*, 29: 537-586.

H. Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press, Cambridge, MA.

Maria Teresa Guasti, Gennaro Chierchia, Stephen Crain, Francesca Foppolo, Andrea Gualmini and Luisa Meroni. 2005. Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20(5): 667-696.

Napoleon Katsos and Dorothy V. M. Bishop. 2011. Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, 120: 67-81.

John Stuart Mill. 1865. *An examination of Sir William Hamilton's philosophy and of the principal philosophical questions discussed in his writings (2nd ed.)*. Longmans, Green and Co., London.

Ira A. Noveck. 2001. When children are more logical than adults: Investigations of scalar implicature. *Cognition*, 78: 165-188.

Anna Papafragou and Julien Musolino. 2003. Scalar implicatures: experiments at the semantics/pragmatics interface. *Cognition*, 86: 253-282.

Bob van Tiel, Emiel van Miltenburg, Natalia Zevakhina and Bart Geurts. In prep. Scalar diversity.

Appendix A. Materials, including variant numbers used

Materials used in this experiment have been extracted from the British National Corpus, distributed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

1. These are supplemented by more than 60/58 massive volumes of press-cuttings. (BNK 829)
2. We just hit at the right moment and from that week onwards, at least 93/90 people turned up. (AB5 566)
3. You may also have noticed that there are more than four grey shades used. (CGU 967)
4. They have lured or chased more than 77/70 species of vertebrates around racetracks in the Kenyan desert, up treadmills at the field station, and over runways of force plates in Milan, all in the interests of learning, as it were, how many kilometres each model gets per litre. (B75 1009)
5. In December 1984 at least 80/86 Jehovah's Witnesses were arrested in Limbé, southwest Cameroon, after holding an unauthorized religious meeting. (A03 628)
6. Violence was believed to be declining; the last war involving more than two great powers had been fought in the Crimea, far away, and the assumptions which governed fighting were more humane than ever before. (CM6 1021)
7. Plant experts at the meeting of the Convention on International Trade in Endangered Species (CITES) have agreed that more than 86/80 species of ‘slipper’ orchids — the genus paphiopedilum from Asia and the genus thragmipedium from South America — should be listed on the CITES Appendix I, which bans all commercial trade. (A59 421)
8. At least 70/77 alternatives have been submitted, with that of "Polish Socialist Labour Party" the front-runner. (A7V 300)
9. In the next example the character's thought spans more than one sentence. (EF8 1488)
10. Iranian-born Khoei, a scholar who had written more than 90/93 books on theology, was known for his adamantly apolitical stance. (HLN 2053)
11. On the basis of earlier work relying on measuring footprints, it had been estimated that there must be at least 58/60 rhinos in the park. (J3K 92)
12. We only have to look at Tintswalo Hospital (Gazankulu) and more than three surrounding villages that fall under the jurisdiction of Lebowa Authority for evidence of this inaccessibility. (FBH 1174)

Appendix B. Mean ratings by-sentence

Tables 2 and 3 present the mean ratings for each sentence in versions 1 and 2 of the experiment. Sentences are numbered as in Appendix A; where applicable, the first-given number in Appendix A was used in version 1 of the experiment, and the second-given number was used in version 2 of the experiment.

Sentence	(i)	(ii)	(iii)	(iv)
1	2.94	3.29	4.06	3.06
2	3.59	3.88	2.94	3.53
3	1.82	3.65	3.82	3.71
4	3.18	3.88	3.18	3.47
5	3.59	3.59	3.76	3.29
6	1.88	3.53	3.59	3.82
7	3.76	3.71	3.35	3.35
8	3.24	3.59	3.94	3.29
9	2.06	3.29	3.18	3.29
10	3.47	3.65	4.29	3.00
11	2.59	4.00	3.29	3.47
12	2.06	3.71	3.29	3.53

Table 2: Mean results by-sentence in version 1 of the experiment

Sentence	(i)	(ii)	(iii)	(iv)
1	4.07	3.57	3.29	3.14
2	3.14	3.64	4.00	3.07
3	1.86	3.36	3.43	3.79
4	3.46	3.50	3.93	3.07
5	3.79	3.93	3.07	2.93
6	2.43	3.21	2.79	3.64
7	4.07	3.29	4.00	2.79
8	3.21	3.64	3.57	3.07
9	2.14	3.14	2.93	3.64
10	3.57	3.50	3.36	2.36
11	3.50	3.92	3.93	2.64
12	2.00	3.43	3.14	3.21

Table 3: Mean results by-sentence in version 2 of the experiment

Priming and Alignment of Frame of Reference in Situated Conversation

Simon Dobnik[1], **John D. Kelleher**[2] **and Christos Koniaris**[1]
[1]University of Gothenburg, Centre for Language Technology,
Dept. of Philosophy, Linguistics & Theory of Science, Gothenburg, Sweden
[2]Dublin Institute of Technology, Applied Intelligence Research Centre,
School of Computing, Dublin, Ireland
{simon.dobnik,christos.koniaris}@gu.se, john.d.kelleher@dit.ie

Abstract

In this paper, we study how the frame of reference (FoR) or perspective is communicated in dialogue through mechanisms such as linguistic priming and alignment (Pickering and Garrod, 2004). In order to isolate the contribution of these mechanisms we deliberately work with a constrained artificial dialogue scenario. First we collect data that deal with human behaviour in interpreting descriptions that are ambiguous in terms of the FoR. From these interpretations we extract and identify strategies for FoR assignment in conversations which we then apply to generate descriptions and measure human agreement with the system. Our findings confirm that both speakers and hearers rely on such mechanisms in conversation.

1 Introduction

A necessary basis for a successful human-machine interaction in a situated dialogue is the ability of the machine to understand and generate spatial references to objects in the spatio-temporal and discourse contexts. Studies of human-human communication, e.g., (Levelt, 1989), reveal that the speaker often uses projective spatial descriptions, e.g., "to the left of the chair" or "in front of the chair" without explicitly specifying the frame of reference, or perspective, according to which the hearer should interpret a scene. In principle, these spatial descriptions may be interpreted *relative* to either of the conversational participants ("...from my perspective", "...from your perspective") or to any other individual or object in the scene ("...from sofa's/Alex's position"). In order to be able to set the orientation of the coordinate frame such objects must have identifiable front and back. We avoid describing FoR as speaker-relative and hearer-relative as in a conversation their roles may change. Instead we refer to system-relative (S) and human-relative (H) FoR. Finally, the FoR may also be assigned *intrinsically* by the landmark/reference object ("the chair") (Levinson, 2003) which we mark as I.

Our long term research goal is to create artificial conversational agents that can participate in situated dialogue. Such an agent must be able to understand and use locative expressions, including those that are dependent on FoR. The agent must resolve the FoR before a geometric spatial template, representing, for example, a region corresponding to "to the left of", can be applied as the FoR sets the origin and the orientation of the coordinate system in which the spatial template is projected (Maillat, 2003). Possibly the simplest approach to handling the FoR issue that can be adopted when creating an artificial conversational agent is to assume or require that all FoR usage is relative to the artificial agents perspective. Unfortunately, however, our earlier work with a situated robot (Dobnik, 2009) shows that relativising all human spatial descriptions to the perspective of the robot adds considerable noise to the data which affects the performance of classifiers that attempt to capture spatial templates. Trafton et al. (2005) show that robots capable of making perspective shifts are more effective in interpreting human descriptions and Steels and Loetzsch (2009) show that they are more successful in learning and generating situated language. However, both approaches do not equip the robots with a model of perspective of the most likely FoR their conversational partner would expect which is the focus of our current study.

There are a number of factors that affect the choice FoR, including: task (Tversky, 1991), personal style (Levelt, 1982), arrangement of the scene and the position of the agent (Carlson-Radvansky and Logan, 1997; Kelleher and

Costello, 2009; Li et al., 2011), and the presence of a social partner (Duran et al., 2011). In this work, however, we focus on *linguistic priming* and *alignment*. By "linguistic" we mean expression of and exposure to content of linguistic utterances. We use the term linguistic priming to distinguish it from and relate it to other forms of priming, for example visual priming by the visual properties of the scene, and priming by the participant role in conversation (speaker/hearer). By alignment we mean adoption of common patterns of behaviour. Watson et al. (2004) conduct psychological studies that confirm the alignment of FoR between conversational partners following a linguistic priming. Johannsen and de Ruiter (2013) investigate further whether the alignment is due to priming or due to preference for a particular FoR in conversation and conclude that there is an interplay of both factors. In contrast to (Watson et al., 2004) and (Johannsen and de Ruiter, 2013) we designed a more complex structure of dialogue games where, for example, a priming step is followed by two interpretive steps before switching the communicative roles of participants, which allows us to study the attenuation of priming and the development of alignment.

Our study includes two experiments which were performed in a constrained spatial environment and dialogue (i) to control the influence of other non-linguistic priming factors, and (ii) to test how humans assign FoR at those points in dialogue where the FoR assignment is at stake: directly after a priming utterance, dialogue turns following this turn and subsequent dialogue turns where the interlocutors switch their roles (from interpretation to generation and vice versa). By examining the behaviour of dialogue participants at these dialogue points we address the following research questions: (i) do participants align their FoR with the linguistically primed FoR used by their dialogue partner; (ii) does the effect of priming degrade over dialogue games; and (iii) does priming persist over role changes?

Overall, if priming develops into alignment, it shows that agents behave cooperatively to their conversational partner (Clark and Wilkes-Gibbs, 1986). In dialogue each conversational participant has a dialogue game-board which contains their individual representation of the state of the dialogue (Ginzburg and Fernández, 2010). One part of the dialogue game-board is the common ground which contains assumptions that conversational participants believe that they have agreed upon. In the priming game (which contains an unambiguous utterance relative to the visual scene) both the hearer and the speaker push the FoR from the speaker's utterance to their common ground; the speaker when they choose what to describe and the hearer when they confirm that they have understood the utterance. In the subsequent ambiguous games both agents have a choice: should they generate and interpret the utterance relative to the FoR that is in the common ground of their dialogue game-board or should they update the FoR in their common ground with a different one. We hypothesise that if the agents are cooperative, they will tend to minimise the updates to the common ground unless this is not necessary, for example, there is no new priming of the FoR through other priming factors. We interpret the non-variability of the FoR in the common ground as alignment. Note that our notion of alignment is slightly different from (Watson et al., 2004) and (Johannsen and de Ruiter, 2013) who consider alignment to occur if a hearer primed with a particular FoR would use this FoR in their next utterance as a speaker. In our framework, alignment occurs earlier, at the point after the hearer updates their common ground with the primed utterance.

Our experiments study the dynamics of FoR updates to common ground in a restricted scenario. In Experiment I the system has no knowledge of the strategies for FoR assignment, instead we try to capture them through observing the behaviour of a human. The system primes the human with an unambiguous scene description and we capture what a human would do in terms of FoR assignment in the subsequent conversational games over visually ambiguous scenes, first when they have a role of the interpreter and finally when they become a generator. In Experiment II we test whether the human strategies for assigning FoR from Experiment I can be used by the system and whether human observers evaluate such behaviour positively. Here, the human primes the system in the first conversational game and in the subsequent games the system has a role of the generator and finally an interpreter of visually unambiguous scenes.

2 Experiment I: alignment of FoR

The focus of the reported research is to investigate the role of linguistic priming and alignment

in FoR-usage in constrained situated dialogues in order to discover an inventory of strategies that an intelligent virtual agent could use to generate and interpret FoR-dependent locative expressions correctly. As a basis for our analysis of strategies, we collected a dataset of situated dialogues. To collect the data, we created a virtual scene embedded in a web-page in which a pre-scripted agent interacts with a human through a series of utterances in particular spatial scene configurations as shown in Figure 1.

Katie: I chose the blue box to the left of the chair.

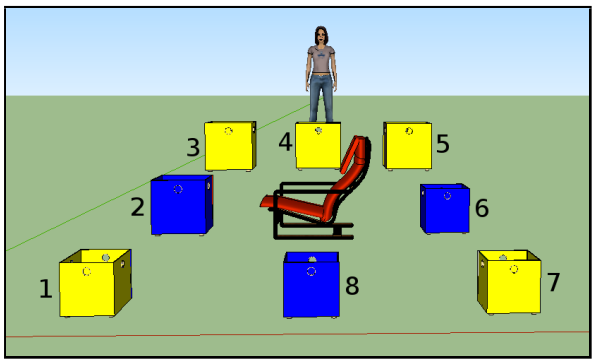


Figure 1: A scene from the virtual environment as seen by a human (not including numbers). The system (the character Katie) generates a description for which the human should decide on the most likely FoR by clicking on a box (2 = H*uman*, 6 = S*ystem* and 8 = I*ntrinsic*).

Conversational Games I

1. The system primes a human for FoR unambiguously: the scene contains only one blue box.
   S: "I chose the blue box to the left of the chair."
   H: Clicks on the intended target object.
2. The system generates an ambiguous description: the scene contains 3 blue boxes, one for each FoR interpretation (cf. Figure 1).
   S: "I chose the blue box to the left of the chair."
   H: Clicks on the intended target object.
3. Identical to Game 2 but with a different spatial description ("to the right of") and a different arrangement of blue boxes.
4. The system asks a human to describe the object that it chose (and marked by an arrow).
   S: "Tell me: which box did you choose?"
   H: Types in their description.

We deliberately opt for such a constrained artificial scenario for two reasons arising from our previous work where we examined assignment of FoR in unrestricted conversation between humans (Dobnik, 2012). Firstly, even if a dialogue task is designed to maximise the usage of spatial descriptions, for example as a variant of the map task (Anderson et al., 1991), longer sequences of potentially ambiguous utterances in respect of the FoR assignment are in minority and therefore one would need to collect a several times larger corpus to obtain a representative number of examples. Secondly, previous studies have shown the FoR assignment is influenced by several factors (task, arrangement of the scene, position of the agent and presence of the social partner) and hence a constrained scenario may be to our advantage as these factors can be controlled. In this study it is not our intention to model human dialogue as a whole but to extract the strategies of FoR assignment through linguistic priming at particular points of dialogue where its assignment is at stake in such a way that the strategies can be used for assignment or disambiguation of FoR in a dialogue manager.

We represent these points in dialogue as a sequence of four dialogue games (each consisting of two turns) which we summarise under the heading Conversational Games I. The conversation was initiated by the system in what we call the priming step (Game 1). This was followed by three games which were intended to show the development of linguistic priming into an alignment of the other agent, the human. Game 2 tested the effectiveness of FoR priming, Game 3 tested the persistence of priming under the same speaker-hearer roles and Game 4 tested the persistence of priming if the speaker-hearer roles change. The system had no knowledge about the FoR assignment (human (H), system (S) or intrinsic (I), i.e., relative to the chair). Rather, the study was intended to capture what FoR an interpreter and finally a generator of an utterance would assume after being linguistically primed for a particular FoR.

Data were collected from both supervised lab sessions and anonymous online contributions. In both cases the same web-interface was used. In total there were 75 trials from which 51 were completed and used in the study. Each participant made judgements for 12 games in total, i.e., 4 games for each of the 3 primed FoRs. All subjects were primed for FoR in the same order which was H $>$ I $>$ S. Table 1 shows conditional probabilities of a human selecting a particular FoR in each subsequent dialogue game following linguistic priming in Game 1. They reveal that priming in

Game 1 does have a strong effect on the human's choice of FoR in the subsequent games (the highest probabilities for each game given each priming are emphasised). Generally, humans align to all 3 FoR primed by the system in Game 2 and to H and I in Games 3 and 4. In Games 3 and 4 the alignment to S loses to the preference for I. This indicates that priming to H and I is persistent in conversation over several games but not priming to S the use of which persistently drops across subsequent games. The priming to H and I also carries over to the fourth conversational game where the speaker-hearer roles change. In more detail, the transition from Game 2 to 3 shows that the alignment to the primed FoR weakens for H and S but it grows stronger for I as shown by the spread of probabilities. This means that as the conversation proceeds there is more variation in the choice of S and H and less in the choice of I. This is because in each game following Game 1 the chosen FoR also adds secondary priming for the following game. If this FoR is the same as in Game 1, it will further strengthen the alignment to the primed FoR, otherwise it will weaken it. In Game 4 where roles change, i.e., human becomes a speaker and system becomes a hearer, an increase in the preference for H and a decrease in the preference for S relative to the previous game is found. This may be because at this stage priming by the speaker role for H is introduced (speakers being egocentric) which competes with the linguistic priming. Overall, at the end of the conversation (Game 4) the perspective that decreases the most is S and the one that remains the most dominant of all three is I.

We explain the increased preference for I at the expense of S if priming was followed on the grounds of the visual priming introduced by the chair. This is more visually salient than the system avatar. It is placed in the middle of the room, appears closer and larger to the human and is red. On the other hand the system avatar is a static character and therefore may lack the salience of an animate person speaking. Given this salience imbalance, humans performing the task may simply forget that they are talking to an agent and consequently focus on the chair. We hypothesise that this is the main reason why the usage of S is in decline in Games 3 and 4, although note that at the beginning of the conversation in Game 2 the likelihood of S following a primed S is higher than H following a primed H. Furthermore, the chair is also a convenient compromise to ground the FoR in for both the system and a human as it is not one of the agents speaking. Visual priming of the chair is constant throughout the conversation whereas speaker-related priming changes from one agent to another.

	Followed by		
Primed by	H	S	I
	Game 1		
H	1.000	0.000	0.000
S	0.000	1.000	0.000
I	0.000	0.000	1.000
	$\chi^2(4) = 388, p < 2.2 \times 10^{-16}$		
	Game 2		
H	**0.513**	0.145	0.342
S	0.073	**0.564**	0.364
I	0.098	0.131	**0.771**
	$\chi^2(4) = 75.250, p = 1.764 \times 10^{-15}$		
	Game 3		
H	**0.460**	0.108	0.432
S	0.111	0.426	**0.463**
I	0.083	0.117	**0.800**
	$\chi^2(4) = 52.828, p = 9.256 \times 10^{-11}$		
	Game 4		
H	**0.508**	0.127	0.365
S	0.308	0.250	**0.442**
I	0.175	0.018	**0.807**
	$\chi^2(4) = 33.613, p = 8.945 \times 10^{-7}$		

Table 1: The probabilities of selecting a particular FoR for each subsequent game given some priming (Game 1). The system primes all FoRs equally and the figures show that all participants correctly identified the unambiguous target object. The χ^2-test confirms the statistical significance of the differences in observed assignments/probabilities. We calculate the χ^2 statistic for each game separately which ensures independence of observations in respect to individuals.

Table 1 shows us whether linguistic priming of FoR initiated by the system in equal proportions develops into alignment of a human. Unfortunately, for this reason we are not able to extract the preference of humans for FoR in the priming Game 1. This would tell us the overall preference for FoR in this spatial and dialogue contexts in the absence of linguistic priming. We estimate this preference in Experiment II in Section 4.

3 Strategies of FoR assignment

How can the strategies for FoR assignment discussed in the previous section be integrated within a dialogue manager of a conversational agent? One way of representing them is using a simple 4-state graphical model as shown in Figure 2, where each state represents a dialogue game and contains a conditional probability table representing the likelihood of the chosen FoR (H, I or S) in that game, given that a particular FoR was chosen in the previous game. The graphical model can be

applied as a classifier within dialogue rules that update the dialogue game-board.

Table 2 shows the conditional probabilities tables for states G_2, G_3 and G_4 of the network. If we choose maximum a posteriori hypothesis, the most likely choice of a FoR for a dialogue manager is always the same FoR as in the preceding step, except at the switch of the conversational roles in Game 4 where S chosen in Game 3 is followed by H in Game 4. Hence, due to the strong alignment of subjects in our experimental scenario the FoR assignment could be implemented in a dialogue manager with only two rules: **If** you are changing your role from interpreter to generator **and** the last FoR was grounded in the location of your conversational partner, **then** ground the FoR in your location; **else** do nothing.

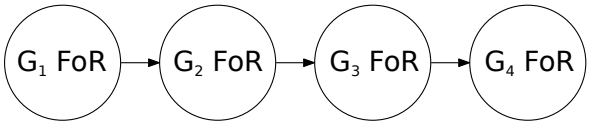


Figure 2: Block diagram of the Bayesian network. Each state of the network represents the model of FoR assignment for a particular dialogue game ($G_1 \ldots G_4$), a sequence of generative and interpretive turns.

	Current game		
Previous game	H	S	I
	Game 1 Priming		
	Game 2		
H	**0.513**	0.145	0.342
S	0.073	**0.564**	0.364
I	0.098	0.131	**0.771**
	Game 3		
H	**0.792**	0.021	0.188
S	0.128	**0.766**	0.106
I	0.011	0.011	**0.979**
	Game 4		
H	**0.833**	0.119	0.048
S	**0.515**	0.364	0.121
I	0.064	0.021	**0.915**

Table 2: The conditional probabilities of selecting a particular FoR in the current game given a particular FoR in the previous game.

4 Experiment II: Application of the FoR alignment

In Experiment I we have shown how humans align their interpretation and generation of utterances involving FoRs to the linguistic priming by the system. We can now use the strategies of human alignment in the system to predict the most likely FoR for the utterance in a dialogue after the system has been primed by the human. In Experiment II we examine whether humans agree with the system using these strategies. In particular, would a human choose the same FoR as the system when it is generating unambiguous descriptions in Games 2–3 after being primed by a human in Game 1? Moreover, would a human taking on a speaker role in Game 4 also choose the FoR that the system would predict given the alignment strategies? To answer these questions we tested whether human strategies for interpretation of FoR could be used by the system for generation and vice versa as summarised in Table 3. We hypothesise that in this new scenario our conversational agent is maximally cooperative with its human partner as it is able to predict and foresee their beliefs and thus minimise the differences in their individual common grounds which would lead to misunderstandings. Hence, we expect that humans interacting with the system will evaluate its performance favourably.

Scenario	Games 1–3	Game 4
Experiment I	interpretation	generation
Experiment II	generation	interpretation

Table 3: The application of the FoR strategies in each experiment.

The listing Conversational Games II summarises the dialogues from Experiment II. In Game 1 the human is invited to prime the system. In Games 2 and Games 3 the human is first offered to choose an object whose location should be described, i.e. a box, then the system generates an unambiguous description of the box using the alignment model and asks the human for agreement. The human can acknowledge their agreement or provide a corrective description. We let humans choose the target box themselves as this gives them the opportunity to build their own representation of the scene before they hear the system's description. This way we attempt to counter the secondary priming introduced by the system's description which may lead human evaluators to overly agree with the system. Game 4 is similar to Games 2 and 3 except that in Game 4 both the human and the system generate a description and the system does so in the background. They agree if they both independently choose the same FoR.

We adapted the web-based environment used in Experiment I to the new scenario. The participants were instructed that they were engaged in a conversation with an artificial agent represented by the character facing them at the opposite side of the room (cf. Figure 1). In order to avoid complex

descriptions such as "the box at the front and to the left of the chair" that are ambiguous between H and I, the corner boxes 1, 3, 5 and 7 were removed from the scene. The scene thus contained only 4 boxes which were all yellow. Humans communicated with the system by choosing a sentence from a list. This was considered appropriate in this context as we are only interested in the alignment of FoR and not in the spontaneous human generation. The sentences differed in respect to the choice of the spatial description and therefore FoR as shown in Game 2 in the listing. The evaluation was performed entirely through online crowdsourcing. Before starting, each participant had to supply a valid email address which attempted to prevent random participation. In total, judgements from 58 complete trials were collected (whereby one participant completed Games 1–3 twice which gave us 59 judgements for these games).

Conversational Games II

1. Human primes the system by describing a focused box.
   - S: "Where is the blue box?"
   - H: "The blue box is {to the left of | in front of | to the right of} the chair."
2. Human chooses a box, the system uses the model for FoR, generates a description and asks the human for agreement.
   - S: "Please choose any box."
   - H: Clicks on one box.
   - S: Using the model and the chosen box: "Aha, you chose the box in front of the chair. Would you agree?"
   - U: "Yes, the box is in front of the chair." | "No, the box is {to the left of | behind} the chair."
3. Identical to Game 2.
4. Human chooses a box which becomes the object in focus. The system asks the human to describe it and makes the assumption about the FoR the human would choose. The exchange succeeds if both are the same.
   - S: "Please choose any box."
   - H: Selects one box by clicking.
   - S: "OK. Now, please tell me: where is the box that you chose?"
   - H: "The box is {to the left of | ...} the chair."
   - S: "Thank you."

4.1 FoR to initialise conversation

In Experiment I the priming of the FoR was a task of the system which assigned the FoR in equal proportions. In Experiment II we want to test how adaptable is the system to the human and hence priming was a task of the human. Their preferences are summarised in Table 4. These probabilities can be used for initialising the conversation (cf. Section 2) and also tell us the preference of humans for FoR in the chosen visual and dialogue contexts; other contexts may lead to different preferences. The figures confirm the general tendencies already described in Section 2. There is a clear hierarchy of the FoR choice to start a conversation, which is I > H > S. However, one confounding factor impacting on this result is the fact that relationship between the FoRs and the spatial descriptions in Game 1 of the evaluation was kept constant across all participants. In particular, I was always associated with describing the blue box as being "in front of" the chair. Several researchers, for example (Logan, 1995; Franklin and Tversky, 1990), have reported results that humans find it easier to use and generate "front" and "back" descriptions rather than "left" and "right". Consequently, this preference for I, although consistent with other research (Kelleher and Costello, 2005; Johannsen and de Ruiter, 2013), may be the result of an interaction with the relative ease of using "front" and "back". In future work we intend to study this confounding factor in more detail.

Game	H	S	I
1	0.4068	0.0508	0.5424

Table 4: The likelihood of human selecting a FoR given the beginning of the conversation.

Moratz and Tenbrink (2006) report that humans prefer to use addressee-centred FoR and therefore adapt to their partner rather than take their own perspective which appears to be contradicted by our results as S is rarely used in comparison to H. When describing scenes humans prefer to use their own perspective over the perspective of the addressee, the system. However, speakers in Experiment II are performing different speech acts than those in (Moratz and Tenbrink, 2006): in the former they are providing a description and in the latter they are issuing a command to a person operating a robot. In (Moratz and Tenbrink, 2006) the hearer of the utterance is much more marked than in Experiment II which may count as a possible explanation for different experimental observations.

4.2 Human agreement with the strategies

As shown in Conversational Games II, in Games 2 and 3 the system used the FoR assignment strate-

gies defined in Section 3 to predict the most likely FoR to generate a description and in Game 4 to make an assumption about the FoR in the description made by its human partner. Table 5 shows a confusion matrix between a system-predicted FoR and a human-chosen FoR. In Games 2 and 3 the human made a corrective description *after* they had heard the system's description. In Game 4 each made their choice independently. The term agreement may be interpreted as a satisfaction of a human with the system's generation in Games 2 and 3 and as a match in their predictions in Game 4. Note that the S is rarely chosen. This is because this FoR was disfavoured by humans in the priming step as shown in Table 4.

		Human		
Game	System	H	S	I
2	H	**22**	0	2
	S	0	**2**	1
	I	0	0	**32**
	Agreement	94.92%		
3	H	**22**	0	2
	S	0	**2**	1
	I	1	0	**31**
	Agreement	93.22%		
4	H	**18**	3	6
	S	0	0	0
	I	0	1	**30**
	Agreement	82.76%		

Table 5: Confusion matrix for the FoR chosen by the system and humans.

Overall, there is a high agreement of humans with the generations of the system: 94.92% in Game 2 and 93.22% in Game 3. The system does slightly less well predicting the FoR assumed for the subsequent generation of a human (82.76%). However, here both were "blind" to each others choice and hence the figure excludes the effect of a potential secondary FoR priming of a human in Games 2 and 3. The system and humans most disagree when the former predicts H but a human chooses S or I. Again, this variability of choice may be explained by the fact that the speaker-hearer roles have reversed and therefore the linguistic alignment is less stable in this new conversational context.

5 Discussion

The results from both experiments show that conversational partners act in a cooperative manner and they align to the linguistically primed perspective. This is the most frequently chosen strategy in this restricted scenario. However, linguistic priming is not the only strategy that they can use for FoR assignment: they may associate FoR with a salient centrally located reference objects (visual priming) or with the speaker or the addressee of the utterance depending on the utterance's speech act (priming by the participant's role in conversation). Both strategies exhibited a secondary effect in our experimental environment.

Directionals are a clear example that the meaning of linguistic expressions is dynamic and consistently changes through updates from the contexts in which the words are used (Larsson, 2007). Applying them in our constrained scenario demonstrates the plasticity of their meaning. An expression like "the box is to the left of the chair" is not only ambiguous in the assignment of the FoR but also in terms of the spatial template projected within the FoR, depending on the arrangement of the scene and the presence of distractor objects (Costello and Kelleher, 2006; Brenner et al., 2007). It follows that the meaning of directionals (and many other kinds of descriptions) relies on both the discourse and perceptual contexts in which they are used. If the meanings of words are dynamic and adaptable to contexts, it must be the case that there exist invariances within the contexts that are stable enough over time to be suitable referents. For example, reference objects in spatial descriptions ("the chair" in the example above) must not change size, shape and location in order to be good landmarks for "the box". The same holds for the discourse context where stability is achieved through alignment. If conversational participants choose the FoR randomly for each utterance, the information that is in the common ground of the dialogue (the sequence of the assigned FoRs) is not a reliable predictor of the forthcoming FoR choices. Participants would have to opt for some other strategy. This would be uncooperative given that linguistic interaction is the primary activity that they are engaged in. Grounding a different FoR in the common ground could also be due to miscommunication (the disagreement in Table 5) which is resolved between participants through alignment (see Mills and Healey (2008)). We hope to study the convergence of participants to a common FoR in case of miscommunication in our future work.

An important question we need to address is how well the strategies that we observe in the constrained scenario generalise to real situated dialogue. There are at least three issues at stake.

In real situated scenes there may be additional invariances in both linguistic and visual contexts that our experimentation did not take into account. This has been addressed extensively in previous research (cf. Section 1) and no doubt will be further investigated. Another question is how these invariances would be used for FoR assignment in cases where all of them are available. Our results suggest that linguistic priming may be stronger than visual priming which may be stronger than speaker priming. For example, the maximum probabilities for selecting each FoR in Game 2 in Table 1 tend to go with the linguistically primed FoR (in a diagonal) rather than visually primed FoR (column I) or speaker primed FoR (column S in Games 1–3 and column H Game 4). It is true that in the subsequent turns the linguistic priming degrades slightly but still has a considerable effect. Notice that in the absence of linguistic priming visual priming takes the lead (Table 4). Thirdly, real conversations may not consist of exactly four conversational games. The states that we explore in our constructed dialogues represent the key transitions between conversational games where the FoR is at stake and the speaker and the hearer must make a choice, namely at the beginning of the conversation, at a continuation of the conversation and at the change of the speaker-hearer roles. Hence, one could apply individual parts of the network to the relevant transitions in a dialogue. Finally, in a real scenario the sequences of conversational games that we explored may be interpreted by intermediate dialogue games that do not involve spatial reasoning. Would linguistic priming degrade in such cases and if so after what length of interruption? Does priming from an intermediary non-spatial dialogue game interfere with priming in a spatial game? This question would have to be answered by further experimental work.

6 Conclusions and future work

We established and tested strategies of perspective taking of conversational participants in a constrained situated dialogue where we focused on linguistic priming. From the collected dataset we can conclude that (i) in the absence of linguistic priming there exist preferences for the assignment of FoR in this scenario, namely *Intrinsic* $>$ *Speaker* $>$ *Hearer* (naming FoR after the conversational roles); (ii) the linguistic priming of FoR at the beginning of a conversation by one participants develops into alignment of both participants in the subsequent games, even when, but to a lesser degree, the speaker-hearer roles change; and (iii) visual properties of scenes and shifts in the speaker-hearer roles also exert priming and consequently affect the alignment to linguistic priming. Through the application of the FoR assignment strategies, we have demonstrated that humans evaluate them favourably, and the properties of the FoR assignment (i–iii) also hold. We additionally demonstrate that a model of interpretative judgements can be used for generating descriptions and vice versa. We expect that the user adaptation of the system would facilitate more effective spatial communication.

We chose a scenario with constrained visual and dialogue contexts to study the strategies of linguistic priming and alignment of FoR with an intention of formulating them as dialogue manager rules. In such a system the FoR assignment model would be part of a larger spatial cognition model which would also include a model for spatial templates and a model of world knowledge for prepositional use. An important part of the investigation would be how to make these models interact with each other aiming at the system to behave in a more cognitively plausible manner. An evaluation of the performance of such a situated agent by human observers would tell us how well the strategies identified in the present work generalise to new and less constrained situations.

Throughout our analysis we have noted how the visual priming of the chair may have drawn the participant's attention to the chair's FoR and that the reverse was the case for the static avatar representing the system. In future studies we will investigate the interaction between object salience and the adoption of FoR. We will also investigate the effects of the description choice between "front"/"back" and "left"/"right" on the FoR assignment by varying the priming from the current front-back dimension for I and the lateral dimension for H and S to the opposite. Overall, varying the parameters of the linguistic and visual contexts reminds us of an important theoretical insight that the meaning of linguistic descriptions is highly dynamic and context relative.

Acknowledgements

The authors wish to thank all participants in experiments and three anonymous DialWatt reviewers.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Michael Brenner, Nick Hawes, John D. Kelleher, and Jeremy L. Wyatt. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *IJCAI 2007*, pages 2072–2077.

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

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1): 1–39.

Fintan J. Costello and John D. Kelleher. 2006. Spatial prepositions in context: the semantics of near in the presence of distractor objects. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, Prepositions '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom, September 4.

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 other's perspective despite cognitive cost. *Cognition*, 121(1): 22–40.

Nancy Franklin and Barbara Tversky. 1990. Searching imagined environments. *Journal of Experimental Psychology: General*, 119(1):63–76.

Jonathan Ginzburg and Raquel Fernández. 2010. Computational models of dialogue. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The handbook of computational linguistics and natural language processing*, Blackwell handbooks in linguistics, pages 429–481. Wiley-Blackwell, Chichester, United Kingdom.

Katrin Johannsen and Jan de Ruiter. 2013. Reference frame selection in dialogue: priming or preference? *Frontiers in Human Neuroscience*, 7(667):1–10.

John D. Kelleher and Fintan J. Costello. 2005. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-SIGSEM workshop on the linguistic dimensions of prepositions and their use in computational linguistics formalisms and applications*, pages 119–127, University of Essex, Colchester, United Kingdom. Association for Computational Linguistics.

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

Staffan Larsson. 2007. A general framework for semantic plasticity and negotiation. In *Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7)*, Tilburg, The Netherlands.

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

Willem J. M. Levelt. 1989. *Speaking: from intention to articulation*. MIT Press, Cambridge, Mass.

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

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

Gordon D. Logan. 1995. Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology*, 28(2):103–174.

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

Gregory J. Mills and Patrick G. T. Healey. 2008. Semantic negotiation in dialogue: the mechanisms of alignment. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 46–53. Association for Computational Linguistics.

Reinhard Moratz and Thora Tenbrink. 2006. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, 6(1):63–107.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John. A. Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press.

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.

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

Matthew E. Watson, Martin J. Pickering, and Holly P. Branigan. 2004. Alignment of reference frames in dialogue. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, USA.

How domain-general can we be? Learning incremental Dialogue Systems without Dialogue Acts

Arash Eshghi
Interaction Lab
Heriot-Watt University
Edinburgh EH14 4AS
a.eshghi@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
Edinburgh EH14 4AS
o.lemon@hw.ac.uk

Abstract

Dialogue is domain-specific, in that the communicative import of utterances is severely underdetermined in the absence of a specific domain of language use. This has lead dialogue system developers to use various techniques to map dialogue utterances onto hand-crafted, highly domain-specific Dialogue Act (DA) representations, leading to systems which lack generality and do not easily scale or transfer to new domains. Here we first propose a new method which avoids the use of DAs altogether by combining an open-domain, incremental, semantic NL grammar for dialogue - Dynamic Syntax - with machine learning techniques for optimisation of dialogue management and utterance generation. We then focus on a key sub-problem associated with this vision: automatically *grounding* domain-general semantic representations in the non-linguistic actions used in specific dialogue domains. Similar to some recent work on open-domain question answering, we present an algorithm that clusters domain-general semantic representations of dialogue utterances based on computing *pragmatic synonymy*, in effect automatically inducing a more coarse-grained domain-specific semantic ontology than that encoded by open-domain semantic grammars.

1 Introduction

> "How many kinds of sentence are there? Say assertion, question, command? – there are countless kinds: countless different kinds of use of what we call "symbols", "words", "sentences". And this multiplicity is not something fixed, given once and for all; but new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten." (Wittgenstein, 1953)

Perhaps the most unyielding obstacle in the working out of sufficiently general models of meaning in dialogue is the astonishingly wide and open-ended range of communicative effects that people can achieve with language in different contexts of use. This is not just a matter of structural context-dependence of fragments, ellipsis and anaphora for which there are increasingly general accounts (see e.g. Ginzburg (2012); Kempson et al. (forthcoming); Kamp&Reyle (1993)). Even when a fully specified semantic representation in some logical language is derived for an utterance, the communicative import of the representation is severely underdetermined in the absence of a known activity, a 'language-game', that the representation is deployed in. Conversely, even within a simple domain, there's a lot of variation in language use that does not ultimately affect the overall communicative goal of the dialogue. For example, in the travel domain, the following dialogues all lead to a context in which A is committed to booking a ticket for B from London to Paris: (a) *A: Where would you like to go? B: Paris, from London*; (b) *A: Where is your destination? B: Paris, A: And your port of departure? B: London.* (c) *B: I need to get to Paris from London, A: Sure*. These dialogues can be said to be *pragmatically synonymous modulo the travel domain*. What is striking about these simple examples is that much of this synonymy breaks down if one moves to another domain (e.g. example (b) where A is an immigration officer): pragmatic synonymy relations are domain-specific.

To bypass this difficulty, Spoken Dialogue Systems (SDS) designers/researchers have used hand-crafted representations of the communicative content of utterances in specific domains, in the form of Dialogue Acts (DA)[1], designed to capture the

[1]Here we use the term "dialogue act" to encompass the whole semantic representation used, ie. standard dialogue acts such as "inform" together with content such as "desti-

specific information needed to complete specific tasks. DAs operate at the interfaces between the core system components in a SDS - Dialogue Management (DM), Natural Language Generation (NLG), and Spoken Language Understanding (SLU) - and have thus lead to systems that lack generality, and are difficult or impossible to transfer to new domains. DAs form a bottleneck representation between SLU and DM, and between DM and NLG. In addition, from a machine-learning point of view DA representations may either under- or over-estimate the features required for learning good DM and/or NLG policies for a domain.

1.1 Structure of the paper

In this paper, we first propose a novel architecture for data-driven learning of fully incremental dialogue systems with little supervision beyond raw dialogue transcripts, which avoids the use of DAs altogether. DAs are instead generated as emergent properties of semantic representations of utterances in specific domains, formed by combining basic semantic units which are delivered by open-domain incremental, semantic grammars[2].

While we do not dispute people's sensitivity to DAs as more coarse-grained units of meaning, here we operate under the assumption that, given a set, stable domain of language use - such as buying a drink at a bar, ordering food in a restaurant, booking a flight, etc. - to which interlocutors are already attuned, the low-level semantic features of utterances are sufficient to encode their pragmatic force, and therefore, that Dialogue Acts need not be explicitly represented[3].

Instead, the appropriate level of meaning representation for a domain will be learned - rather than hand-crafted/designed - from a set of successful in-domain dialogues with no DA annotations. These dialogues are first parsed using Dynamic Syntax (Kempson et al., 2001; Cann et al., 2005), which maps them to open-domain semantic representations of the final contexts reached by the interlocutors, i.e. the semantic content that they jointly commit to. In order to capture the domain-specific pragmatic synonymy relations described above, we will assume a weak form of supervision: that the dialogues are annotated with representations of the non-linguistic actions taken and when, e.g. a data-base query, a flight booking, serving a drink, etc. A function is then learnt which maps these contexts to the non-linguistic action representations. Effectively, this function maps the very fine-grained semantic ontology encoded by the open-domain DS grammars (or any open-domain semantic parser), onto a more coarse-grained ontology with fewer semantic distinctions, based on pragmatic synonymy. It is an algorithm for learning this function that we then focus on in this paper.

First we review some recent related work, in section 2. Then we present the overall model and framework that we are developing fror this problem, in section 3. In section 4 we present the algorithm we have developed for computing the pragmatic synonymy function.

2 Related work

There has been a recent surge of interest in domain-general or "open-domain" semantic parsing. Most similar to our work is perhaps that of (Allen et al., 2007; Dzikovska et al., 2008) who devise a system for mapping open-domain logical forms in a formalism that is similar to Minimal Recursion Semantics (the LF representation), onto domain-specific representations suitable for reasoning and planning within a specific dialogue domain (the KR representation). However, unlike the architecture proposed here, the ontology mappings are defined by hand, rather than learned from data, and the grammar employed is not incremental.

There's also the work of (Kwiatkowski et al., 2013), who map open-domain CCG semantic parses to Freebase for question-answering. Here, an open-domain Question-Answering system (note: not a full dialogue system) is learned by using a wide-coverage CGG parser over questions. Kwiatkowski et al. (2013) develop a method for automatically mapping CCG semantic LFs onto the Freebase ontology, which is similar in spirit to the algorithm we present in section 4. In our case, the ontology is not that provided by Freebase (although nothing prohibits this), but instead the ontology of back-end application actions used in spe-

nation=Dublin").

[2]Note that these grammars will also deliver generic speech act representations such as "question" and "acknowledgement" which we will learn the import of in specific domains of usage.

[3]The question of how interlocutors come to coordinate on the structure of an activity, i.e. how language-games emerge in the first place, is a challenging one. We put this problem on one side here, but see e.g. Healey (2008); Mills (2013 in press); Mills & Gregoromichelaki (2010).

cific dialogue systems (e.g. searching for a flight from X to Y, paying a bill, etc). At a high level, the problem is similar: mapping domain-general semantic representations onto an ontology, though Kwiatkowski et al. (2013) do not need to consider sequences of sentences / utterances, or dialogue acts. Similar work is presented by (Cai and Yates, 2013b; Cai and Yates, 2013a), who also work using Freebase and do not consider dialogues. Their system maps English words onto individual Freebase symbols, and does not handle conjunctions and disjunctions of ontology symbols, as our approach and that of Kwiatkowski et al. (2013) do.

3 Overall model

Before presenting our main algorithm, we first outline the overall method we propose of combining (1) Dynamic Syntax (DS), a domain-general incremental, semantic grammar framework, shown to be uniquely well-placed in capturing the fragmentary and context-dependent nature of spontaneous dialogue (Gargett et al., 2009; Gregoromichelaki et al., 2009); and (2) statistical machine learning with data-driven optimisation methods which are known to robustly handle noise and uncertainty in spoken language. DS will provide the domain-general semantic parsing (i.e. SLU) and surface realisation (i.e. low-level language generation) components, and machine learning for DM will provide the crucial bridge between them and higher-level action and content selection processes. In order to integrate these components, and to use dialogue data for training, we require a 'pragmatic synonymy" function mapping semantic representations provided by DS into specific dialogue system domain ontologies. We present this is section 4.

We first introduce and motivate the particular open-domain semantic parsing formalism that we will use in this work, and then explain the the proposed overall method (see section 3.2).

3.1 Dynamic Syntax and TTR (DS-TTR)

For the required semantic parser, we use a well-established semantic parsing framework, Dynamic Syntax (DS, (Kempson et al., 2001)), which models dialogue as a word-by-word *incremental, interactive* process of constructing meaning representations, with no intermediary layer of syntactic structure over words. We choose this rather than other possible semantic formalisms (e.g. CCG) because it has been shown to be uniquely well-placed in capturing the inherent fragmentary and context-dependent nature of spontaneous dialogue (Eshghi et al., 2012; Gregoromichelaki et al., 2013 in press; Gargett et al., 2009). Since DS is inherently incremental, and not sentence-based, it enables the word-by-word exploration – babbling – of the space of possible grammatical dialogues and their corresponding contexts within a given domain (see e.g. Fig. 3).

In DS, grammaticality is defined as parsability in context; words are associated with conditional Lexical Actions that monotonically update (partial) *semantic trees*, representing predicate argument structure with new semantic information and/or requirements for information to come; there are also Computational Actions, specifying general logical tree operations (e.g. beta-reduction of daughters), and strategies to adjust context for parsing of subsequent words. DS is bidirectional with generation defined in terms of parsing, and operating over the same meaning representations: a dialogue agent can switch from parser to generator (and vice versa) at any point (subsententially, as well as at sentence boundaries), where the generator starts where the parser finished, i.e. the context for generation will be the (partial) semantic tree derived by the parser so far. Dialogue fragments, including corrections, clarification ellipsis, short answers, adjuncts and continued utterances are all modelled grammar- internally in this way (Gregoromichelaki et al., 2009; Gargett et al., 2009).

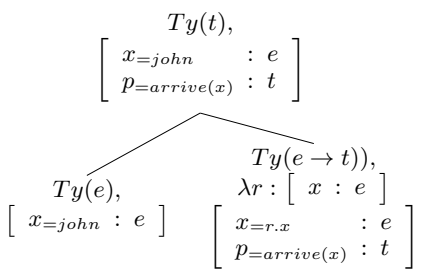


Figure 1: Complete semantic tree for "John arrives". Nodes are decorated with semantic type and formulae.

Type Theory with Records (TTR) Type Theory with Records (TTR) is an extension of standard type theory shown useful in semantics and dialogue modelling (Cooper, 2005; Ginzburg, 2012).

To accommodate dialogue processing, and allow for richer representations of the dialogue con-

text recent work has integrated DS and the TTR framework to replace the logical formalism in which meanings are expressed (Purver et al., 2010; Purver et al., 2011; Eshghi et al., 2012).

In TTR, logical forms are specified as *record types* (RTs), sequences of *fields* of the form $[\, l : T \,]$ containing a label l and a type T. RTs can be witnessed (i.e. judged as true) by *records* of that type, where a record is a sequence of label-value pairs $[l = v]$, and $[l = v]$ is of type $[\, l : T \,]$ just in case v is of type T.

$$R_1 : \begin{bmatrix} l_1 & : T_1 \\ l_{2=a} & : T_2 \\ l_{3=p(l_2)} & : T_3 \end{bmatrix} \quad R_2 : \begin{bmatrix} l_1 & : T_1 \\ l_2 & : T_{2'} \end{bmatrix} \quad R_3 : []$$

Figure 2: Example TTR record types

Fields can be *manifest*, i.e. given a singleton type e.g. $[\, l : T_a \,]$ where T_a is the type of which only a is a member; here, we write this using the syntactic sugar $[\, l_{=a} : T \,]$. Fields can also be *dependent* on fields preceding them (i.e. higher) in the record type – see R_1 in Figure 2. Importantly for us here, the standard subtyping relation $\sqsubseteq$ can be defined for record types: $R_1 \sqsubseteq R_2$ if for all fields $[\, l : T_2 \,]$ in R_2, R_1 contains $[\, l : T_1 \,]$ where $T_1 \sqsubseteq T_2$. In Figure 2, $R_1 \sqsubseteq R_2$ if $T_2 \sqsubseteq T_{2'}$, and both R_1 and R_2 are subtypes of R_3.

3.2 Proposed Overall Method: BABBLE

We start with two resources: a) a wide-coverage Dynamic Syntax parser L (either learned from data (Eshghi et al., 2013), or constructed by hand), for incremental spoken language understanding; b) a set D of transcribed successful example dialogues in the target application domain. Overall, we then need to perform 2 main steps: 1) extract the dialogue goal states from D using L, and 2) automatically generate jointly optimised Dialogue Manager and NLG components.

We then carry out the following steps, explained in greater detail below) to achieve steps 1 and 2:

Step 1.1 Parse all $d \in D$ using L, generating a set of final dialogue contexts, C, each a TTR Record Type representing the grounded semantic content for d; see Fig. 3[4] Collect the successful dialogues in D and extract the set of goal states A, represented as record types;

Step 1.2 Construct the Generalized Goal Context, GGC: the *maximally specific super-type* (the largest common denominator) of A;

Step 2.1 Automatically construct a Markov Decision Process (MDP) for D (see Fig. 3). Generate the state space S using feature function F defined to extract the semantic features (Record Types) in the GGC (i.e. the state space tracks all and only the semantic types present in the GGC), and compute the transition function T via the set of parsed dialogues, use L as the MDP action set, and define Reward function R as reaching the GGC state while minimising time penalties;

Step 2.2 Solve the generated MDP using Reinforcement Learning methods: train an action selection mechanism, where actions are system utterances of the lexical items $a \in L$, optimised via R. This process has a large action set, but action selection will be bounded via a measure of distance from GGC (see below) and is also constrained by the DS grammar.

The result will be the combined DM and NLG components of a dialogue system for D: i.e. a jointly optimised action selection mechanism for DM and NLG, with L providing the SLU component. Domain extension would then be a matter of adding new data and retraining the system. We now describe each of these steps in further detail.

Inducing the dialogue goal (Step 1). Recall the examples of pragmatic synonymy in dialogue given in the introduction, for example

(a) A: Where would you like to go? B: Paris, from London; (b) B: I would like to go to Paris; A: Sure, where from? B: London; (c) A: Where is your destination? B: Paris A: And your port of departure? B: London. (d) B: I need to get to Paris from London A: Sure. These dialogues can be said to be *Pragmatically Synonymous modulo the travel domain*. The source of this variation is twofold: structural, i.e. syntactic and interactional variation; and lexical-semantic, i.e. variation in the basic semantic ontology employed. While (a) and (b) differ only structurally, and not semantically, they differ from (c) and (d) on both levels.

The aim of this step is to extract automatically from D, a compact, tractable representation of a Generalised Goal Dialogue Context (GGC) that captures – abstracts over – both kinds of variation, and which the RL agent will later be trained to track and achieve in the MDP state space. The GGC thus constructed will allow the RL agent not

[4] In all our example context representations in TTR, information about commitment to content, and who said what is suppressed, but see (Purver et al., 2010) for how they are encoded in TTR.

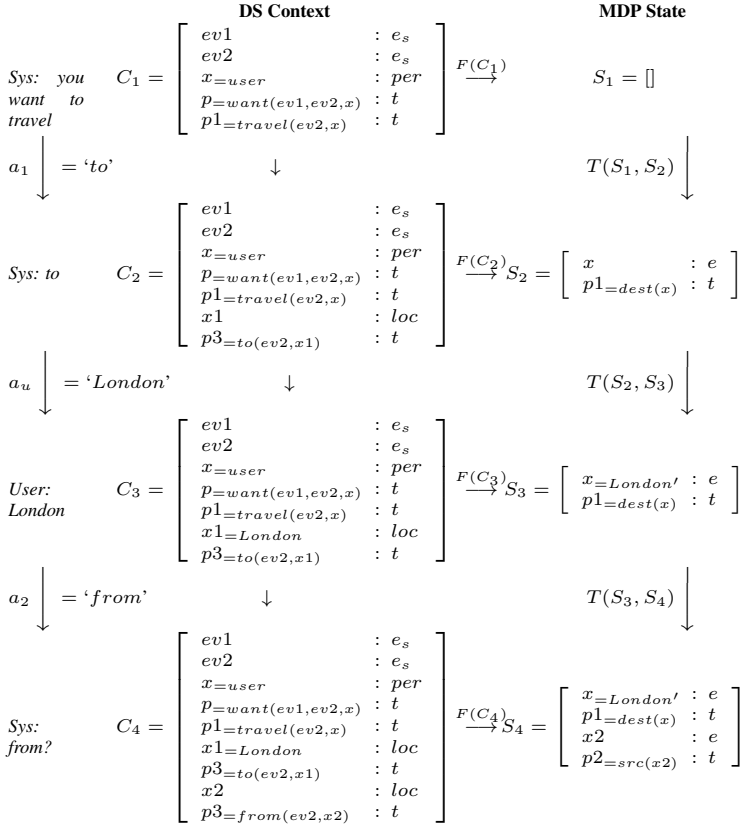


Figure 3: Example incremental action (word) selection via the BABBLE method. See Section 3.2.

only to reproduce the same diversity in its generated utterances, but understand and respond to the diverse language employed by its users, as exemplified in D, without recourse to hand-crafted dialogue act representations. Importantly, the GGC will also serve to constrain the very large space of dialogue policies that the RL agent would otherwise have to search/explore. The construction of the GGC will proceed in two generalisation stages: (1) structural: parsing all the dialogues in D with L producing a set of all the final contexts, C, reached by the dialogues in D; and (2) semantic: partitioning of the set of all semantic features of the C s into a set of equivalence classes, modulo pragmatic synonymy relations, forming, in effect, a domain-specific ontology. We explain these steps below:

1.1. Parsing dialogues with a DS grammar allows us to abstract away from the syntactic and interactional particularities of specific dialogues in D: dialogues are mapped onto domain-general semantic representations of the final contexts jointly established by the interlocutors, in effect allowing us to organise the dialogues in D into a set of structural equivalence classes. For example, dialogues (a) and (b) above will be grouped into the same class in virtue of giving rise to the same final context.

1.2. However, the DS grammar is domain-general, encoding a very fine-grained ontology of semantic types, i.e. lexical variation in the dialogues will always lead to semantic variation in the C's. But much of this variation is pragmatically inconsequential for task success within a given domain: for example modulo the travel domain, dialogues (b), (c) and (d) are pragmatically synonymous (c.f. in the question-answering case, (Kwiatkowski et al., 2013)).

Therefore, our goal here is to *create equivalence classes of the semantic features* (TTR record types) of the C, such that two features are placed in the same equivalence class if they make the same pragmatic contribution to in-domain task success. To achieve this, we can use a weak form of supervision: we can assume that the datasets D contain, in addition to raw dialogue transcripts, representations of the non-linguistic actions taken, e.g. data-base queries, flight bookings, serving a drink; depending on the domain. The semantic features of the C will then be grouped into equivalence classes in virtue of giving rise to the same non-linguistic actions, i.e. in virtue of being pragmatically equivalent. For example, dialogues (b) and (c) above will give rise to different final contexts, but both lead to the same non-linguistic action `book(Source=London, Dest=Paris)`. These action representations encode a domain-specific ontology and provide an interface between the domain-general semantic representations delivered by L and the extralinguistic context of the dialogue task. This process can thus be described as mapping a fine-grained, open domain, semantic ontology onto a more coarse-grained domain-specific one with fewer semantic distinctions, based on pragmatic synonymy relations. The task of finding this mapping is akin to that of (Kwiatkowski et al., 2013) who present a method for doing this, in order to produce an open-domain Question Answering system that uses an open domain CCG semantic parser. This is the main algorithm that we present in section 4.

Other steps are needed, in particular the reinforcement learning of incrementally generating lexical actions so as to achieve the GGC. We leave presentation of this method to future work, and

Mapping Type	Example mapping in FOL (Kwiatkowski et. al)	Example mapping in TTR (this paper)
Collapse (type e)	$\iota.x\ Public(x) \wedge Library(x) \rightarrow PL$	$\begin{bmatrix} r & : & \begin{bmatrix} x & : e \\ p_{=Public(x)} & : t \\ p1_{=Library(x)} & : t \end{bmatrix} \\ x_{=\iota(r.x,r)} & : e \end{bmatrix} \rightarrow \begin{bmatrix} x_{=PL} & : e \end{bmatrix}$
Collapse (type t)	$capital(y) \wedge in(y,x) \rightarrow capitalof(x,y)$	$\begin{bmatrix} x & : e \\ y & : e \\ p_{=capital(y)} & : t \\ p1_{=in(y,x)} & : t \end{bmatrix} \rightarrow \begin{bmatrix} x & : e \\ y & : e \\ p_{=capitalof(x,y)} & : t \end{bmatrix}$
Splitting	$capitalof(x,y) \rightarrow capital(y) \wedge in(y,x)$	$\begin{bmatrix} x & : e \\ y & : e \\ p_{=capitalof(x,y)} & : t \end{bmatrix} \rightarrow \begin{bmatrix} x & : e \\ y & : e \\ p_{=capital(y)} & : t \\ p1_{=in(x,y)} & : t \end{bmatrix}$

Table 1: Examples of the different types of ontology mapping in FOL and TTR

here focus on step 1.2 above.

4 Pragmatic Synonymy: grounding semantic ontologies in action

In this section we describe an algorithm for learning a mapping F from semantic contexts derived from parsing in-domain dialogues with wide-coverage DS grammars, onto representations of the back-end, non-linguistic actions of the system, whose parameters together constitute the MDP state space (see above).

4.1 Types of synonymy mappings

Our aim here can be seen as somewhat similar to the work of Kwiatkowski et al. (2013), where an open-domain Question-Answering system (note: not a full dialogue system) is learned by using a wide-coverage CGG parser over questions. Kwiatkowski et al. (2013) develop a method for automatically mapping CCG semantic parses (of questions, not dialogues) onto a particular knowledge base ontology (in our case, the application back-end actions, such as database searches, flight bookings, etc). Overall, two types of mappings between meaning representations are discussed, *collapsing* and *splitting* ontology constants of different types(e.g. type e or t). Table 1 shows examples of these in First-Order Logic (FOL) as per Kwaitkowski et al. and Record Types (RT) of the Type Theory with Records used in this paper:

As noted by Kwiatkowski et al. (2013), the full set of possible collapses of an input meaning representation MR is limited by its number of constants, since each collapse removes at east one constant. The number of possible collapses is therefore polynomial in the number of constants in MR and exponential in the arity of the most complex type in the ontology. For typical dialogue system domains this arity is only 2 or 3. The splitting operation covers cases where multiple constants in the ontology represent the meaning of a single word. To constrain complexity, we can limit the splitting operation to apply only once for each underspecified constant in MR.

$$\left\langle \begin{bmatrix} ev1 & : e_s \\ ev2 & : e_s \\ x_{=user} & : per \\ p_{=want(ev1,ev2,x)} & : t \\ p1_{=travel(ev2,x)} & : t \\ x1_{=London} & : loc \\ p3_{=to(ev2,x1)} & : t \end{bmatrix}, \begin{bmatrix} x_{=London'} & : e \\ p1_{=dest(x)} & : t \\ act_{=book(x)} & : e \end{bmatrix} \right\rangle$$

Figure 4: Example $\langle C, A \rangle$ pairing. C represents the context reached in: "A: I want to travel to London B: Sure", and A represents a booking action with London as destination

4.2 Problem Statement

Input A set, T, of training examples of the form $\langle C, A \rangle$ where each C is a domain-general record-type (RT) representation of the final semantic context reached by parsing an in-domain dialogue with DS; and A, also a RT, representing the non-linguistic, back-end action taken by the system at the point where C was reached. As such, the A encodes the domain-relevant information required by a dialogue system to complete its tasks. Figure 4 shows one training example in the travel domain.

Output A function $DCont : RecType \rightarrow RecType$ ($DCont$ stands for domain content, and is a function from TTR record types to TTR record types, see section 3.1), determined by a set of ordered pairs, $F = \{\langle c_1, a_1 \rangle, \ldots, \langle c_n, a_n \rangle\}$, which, given new, unseen contexts - but *in part* similar to the training instances - extracts the domain-relevant information from them: F specifies which parts of the semantic information in the contexts - i.e. which supertypes of the context RTs - go on to make up which parts of the target action representations. F determines $DCont$ as follows:
$DCont(x) = \bigwedge_{\langle c,a \rangle \in S} a$, where, $S = \{\langle c, a \rangle \in F | x \sqsubseteq c\}$
($\wedge$ represents the intersection of one or more types

(Cooper, 2005). The intersection is formed by the union of the fields in the record types, with fields that have the same label collapsing into one)
$DCont$ has the following properties:

1. *Many-to-one:* Distinct semantic information in the Cs could, in the general case, be mapped onto the same action representation or parts thereof. This property ensures pragmatic synonymy relations among the supertypes of the Cs. For example, the semantics of "my destination is Paris" and that of "I want to travel to Paris", while being for the most part distinct, will be mapped onto the same booking action in the travel domain.

2. *Surjective over T:* The space of possible target action representations, i.e. the space of the supertypes of the As is fully covered by the mapping. Formally:
$\forall\,(\langle C, A\rangle \in T)\,\exists\,(S \subseteq F)\left[\bigwedge_{\langle c,a\rangle \in S} a = A\right]$

3. *Maximally general over T:*

   (a) $\forall\,(\langle c_j, a_j\rangle \in F)\,\forall\,(\langle C, A\rangle \in T)\,[C \sqsubseteq c_j \rightarrow A \sqsubseteq a_j]$
   i.e. that F generalises to - is correct for - T;
   (b) that anything less specific would not generalise to T:
   $\forall\,(\langle c_j, a_j\rangle \in F)\,\neg\exists c_k$
   $[c_j \sqsubset c_k \wedge \forall\,(\langle C, A\rangle \in T)\,[C \sqsubseteq c_k \rightarrow A \sqsubseteq a_j]]$,
   ensuring that F determines *the minimal* amount of semantic information needed in the contexts to determine some part of an action representation, i.e. that the domain of F remains most general (least specific).
   (c) similarly to (b), that the mappings determine *the maximal* amount of semantic information in the target action representations - the range of F - i.e. that for any $\langle c, a\rangle \in F$ anything *more specific* than a would not be sufficiently encoded by c.

4.3 Learning F

Hypothesising individual mappings using type lattices In processing each training pair $\langle C_i, A_i\rangle$, and enumerating mappings from C_i to A_i, the algorithm makes use of *type lattices*, constructed in advance for all the C_i and A_i. These encode the space of possible super-types of a record type RT - see Fig. 5 - with RT appearing at the bottom node, the empty type [] at the top node, and all super-types of RT in between getting progressively more specified as we move down

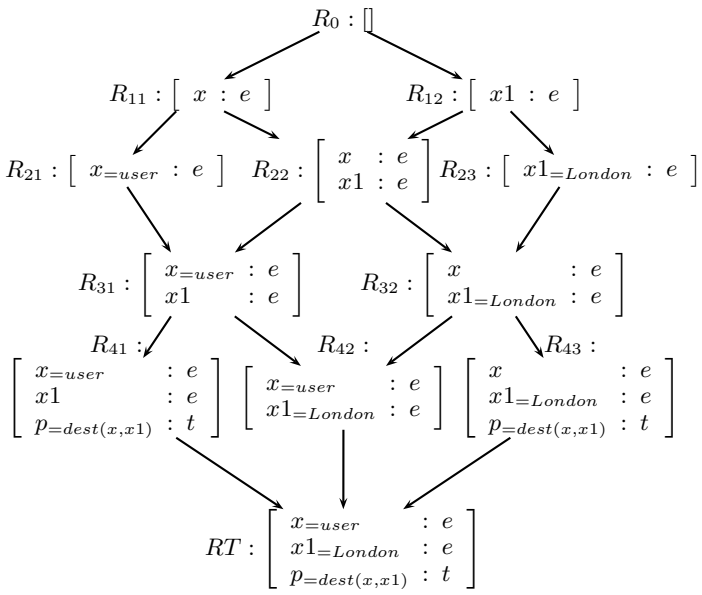


Figure 5: RT hypothesis lattice

the lattice: the lattice is a partial order with $\sqsubseteq$ (is subtype of) being the order relation. Importantly, each edge is also a record type R_I representing the possible *minimal increments* from one RT, R_j, to another, R_{j+1}, such that $R_j \wedge R_I = R_{j+1}$ (see Eshghi et al. (2013) where lattices are similarly used to hypothesise semantic increments in a grammar induction task).

A pair of such lattices for each training example $\langle C_i, A_i\rangle$ (henceforth context lattice and action lattice) thus specifies a partial order on individual mappings $\langle c, a\rangle$ from supertypes of C_i onto supertypes of A_i: we can therefore explore the space of such mappings, in an order that guarantees that the first $\langle c, a\rangle$ encountered, that generalises to other training examples - that satisfies property 3(a) above - also satisfies 3(b) and 3(c), i.e. that c encodes the minimal amount of semantic information needed to determine a: the maximally specific supertype of A_i that generalises. Once any such $\langle c, a\rangle$ is found, we can mark c and a with *pointers* on the lattices, thus partitioning C_i and A_i into what has already been processed/consumed successfully (intersection of RT edges/increments leading to the root above the pointer), and what remains to be processed (intersection of RT edges/increments below the pointers). What remains of the exploration of the space of mappings can now take place, in a recursive fashion, on the sub-lattices whose roots the pointers now mark, with whatever falls outside these sub-lattices ignored in subsequent processing. Furthermore, in the processing of a sub-

sequent training example, $\langle C_j, A_j \rangle$, the mappings already found for previous training examples and stored in F, if applicable (i.e. if for a $\langle c, a \rangle \in F$, $C_j \sqsubseteq c$ and $A_j \sqsubseteq a$) can be 'applied' immediately to $\langle C_j, A_j \rangle$, by moving the pointers on the corresponding lattices to c and a, thus precluding any repetitive processing across the training examples. In fact, given bounded semantic variability within a dialogue domain, if the first few training examples are varied enough, not much will remain to be done for later examples. This process is, in effect, a dynamic programming solution to the problem and thus gives us a handle on its exponential computational complexity.

input : A list T of training pairs $[\langle C_1, A_1 \rangle, \ldots, \langle C_n, A_n \rangle]$
output: The mapping F, a set of ordered pairs

Initialise $F = \{\}$;
Construct/Initialise Lattices for T;
  $lattices \leftarrow [\langle LC_1, LA_1 \rangle, \ldots, \langle LC_n, LA_n \rangle]$;
for $i \leftarrow 1$ **to** n **do**
  $\langle LC, LA \rangle \leftarrow lattices[i]$;
  $\langle LC, LA \rangle$.`MovePointersTo(`*F*`)`;
  while $\neg LA.pointerAtBottom()$ **do**
    CONTEXTINC: **while** `HasMoreIncrements(`*LC*`)` **do**
      $c \leftarrow$ `NextSmallestIncrement(`*LC*`)`;
      ACTIONINC: **while** `HasMoreIncrements(`*LA*`)` **do**
        $a \leftarrow$ `NextLargestIncrement(`*LA*`)`;
        for $j \leftarrow i + 1$ **to** n **do**
          $\langle LC_j, LA_j \rangle \leftarrow lattices[j]$;
          if $C_j \sqsubseteq c \wedge A_j \not\sqsubseteq a$ **then**
            continue ACTIONINC;
          end
        end
        $F.add(\langle c, a \rangle)$;
        $\langle LC, LA \rangle$.`MovePointersTo(`$\langle c, a \rangle$`)`
      end
    end
  end
end

Algorithm 1: Learning F

Details of Algorithm 1 Algorithm 1 details the above process. Given current pointer positions on lattice pairs, $\langle LC, LA \rangle$, the functions, `NextSmallestIncrement(LC)` and `NextLargestIncrement(LA)` return the next least specific, and next most specific increments respectively. These are formed by intersecting the record types corresponding to edges on paths of increasing or decreasing length respectively, downwards through the lattice, from the current pointer position. The implementations of these functions are both in terms of a simple breadth first traversal of the sub-lattices whose roots are marked by the pointers - we suppress any detail here. The `HasMoreIncrements()` function is boolean valued, and determines whether the current sub-lattice is exhausted, i.e. whether all possible increments have already been returned. The function `MovePointersTo(`$\langle c, a \rangle$`)`, applied to a lattice pair, moves the pointers down to c and a on the context and action lattices, as described above. Finally, the inner most `for` loop, checks to see if the current mapping hypothesis generalises to the rest of the training examples, i.e. whether it has the property 3(a) above.

This algorithm covers the mapping types discussed in section 4.1: collapsing and splitting of ontology constants. To further constrain the search, we can incorporate the constraints discussed briefly in that section. Finally, we have not covered functional types here, but TTR affords the full power of the lambda calculus (Cooper, 2005), and these can be incorporated within the algorithm. We leave the details on one side here.

5 Summary and Future Work

We proposed a novel architecture for learning fully incremental dialogue systems with little supervision beyond raw dialogue transcripts and without recourse to dialogue act representations, by combining open-domain, incremental semantic grammars with state-of-the-art machine learning methods for learning NLG/DM policies. We argued that dialogue acts can instead be seen as emergent from learning, and that they need not be explicitly represented. We then focused on a key sub-problem associated with this vision: automatically *grounding* domain-general semantic representations in the non-linguistic actions used in specific dialogue domains. We presented an algorithm for learning such a mapping, which, in effect, clusters parts of domain-general semantic representations of dialogue contexts based on *pragmatic synonymy*, thus inducing a more coarse-grained domain-specific semantic ontology than that encoded by open-domain semantic grammars.

A major part of this paper is a proposal for a programme of research, and hence the most immediate future work consists in carrying out this research and implementing/evaluating the algorithms proposed.

6 Acknowledgments

The research leading to this work was funded by the EC FP7 programme FP7/2011-14 under grant agreement no. 287615 (PARLANCE). We also thank Matthew Purver and Robin Cooper for valuable discussions around the ideas presented here.

References

James Allen, Myroslava Dzikovska, Mehdi Manshadi, and Mary Swift. 2007. Deep linguistic processing for spoken dialogue systems. In *Proceedings of the the ACL workshop on Deep Linguistic Processing*.

Qingqing Cai and Alexander Yates. 2013a. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Qingqing Cai and Alexander Yates. 2013b. Semantic parsing freebase: Towards open-domain semantic parsing. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*.

Ronnie Cann, Ruth Kempson, and Lutz Marten. 2005. *The Dynamics of Language*. Elsevier, Oxford.

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Myroslava O. Dzikovska, James F. Allen, and Mary D. Swift. 2008. Linking semantic and knowledge representations in a multi-domain dialogue system. *Journal of Logic and Computation*, 18:405–430.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper*, pages 325–349.

Arash Eshghi, Julian Hough, and Matthew Purver. 2013. Incremental Grammar Inducion from Child-directed Dialogue Utterances. In *Proc. of ACL Workshop on CMCL*.

Andrew Gargett, Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, and Yo Sato. 2009. Grammar resources for modelling dialogue dynamically. *Cognitive Neurodynamics*, 3(4):347–363.

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

Eleni Gregoromichelaki, Yo Sato, Ruth Kempson, Andrew Gargett, and Christine Howes. 2009. Dialogue modelling and the remit of core grammar. In *Proceedings of IWCS*.

Eleni Gregoromichelaki, Ruth Kempson, Christine Howes, and Arash Eshghi. 2013, in press. On making syntax dynamic: the challenge of compound utterances and the architecture of the grammar. In Ipke Wachsmuth, Jan de Ruiter, Petra Jaecks, and Stefan Kopp, editors, *Alignment in Communication: Towards a New Theory of Communication*. Series: Advances in Interaction Studies (series editors: Kerstin Dautenhahn, Angelo Cangelosi).

Patrick G. T. Healey. 2008. Interactive misalignment: The role of repair in the development of group sublanguages. In R. Cooper and R. Kempson, editors, *Language in Flux*. College Publications.

Hans Kamp and Uwe Reyle. 1993. *From Discourse To Logic*. Kluwer Academic Publishers.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.

Ruth Kempson, Ronnie Cann, Arash Eshghi, Eleni Gregoromichelaki, and Matthew Purver. forthcoming. Ellipsis. In S. Lappin and C. Fox, editors, *Handbook of Contemporary Semantics*. Oxford University Press.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

G. Mills and E. Gregoromichelaki. 2010. Establishing coherence in dialogue: sequentiality, intentions and negotiation. In *Proceedings of SemDial (PozDial)*.

Greg J. Mills. 2013, in press. Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*.

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. 2010. Splitting the 'I's and crossing the 'You's: Context, speech acts and grammar. In P. Łupkowski and M. Purver, editors, *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–50, Poznań, June. Polish Society for Cognitive Science.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proc. Int. Conference on Computational Semantics*, pages 365–369.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell Publishing.

Persuasion in Complex Games

Markus Guhe, Alex Lascarides
University of Edinburgh, School of Informatics
Informatics Forum, 10 Crichton St
Edinburgh EH8 9AB, Scotland
m.guhe@ed.ac.uk, alex@inf.ed.a.uk

Abstract

We study the power of persuasion in a game where each player's own preferences over the negotiation's outcomes are dynamic and uncertain. Our empirical set up supports evaluating individual aspects of the persuasion and reaction strategies in controlled ways. We show how this general method gives rise to domain-specific conclusions, in our case for *The Settlers of Catan*: e.g., the less scope there is for persuading during the game, the more one must ensure one gains an immediate benefit from it beyond the desired trade.

1 Introduction

In this paper we study persuasion in a non-cooperative setting, which Gricean (1975) maxims don't account for (Asher and Lascarides, 2013). Within game theory, standard negotiation models ascribe each player complete and static knowledge of his own (intrinsic) preferences over the negotiation's outcomes (e.g., Binmore (1998)). So the associated models of *persuasion* focus only on the persuader manipulating his opponents' *beliefs* about which outcomes are likely (e.g., Rubinstein (2007)). For instance, during trading, the receiver of an offer to exchange wheat for clay might declare he has no wheat, and indeed be lying, so as to persuade his opponent to accept his counteroffer of ore for clay.

But if trading is a fraction of the action sequence in a complex game, then a player's estimates of which next trade would enhance (or hinder) his chances to eventually win may be wrong. Persuasion then has higher stakes: there's a new potential *payoff* in manipulating an opponent's *preferences* over the next trade, not just his beliefs; but there's also a new *risk* because the persuader's deficient perception of the potential benefits of a particular trade may mean persuading backfires on him. In addition, the persuader risks revealing information about his own intentions or preferences via the persuasion move.

Studies on manipulating an opponent's trading preferences exist in argumentation theory (e.g., Amgoud and Vesic (2014)), but these models focus entirely on the logical structure of successful persuasion moves—i.e., moves where the recipient is persuaded and so changes his behaviour in the intended way. They don't consider the persuading agent having a false perception of his own payoffs, and so don't model the above risk of successfully persuading in a complex game.

Persuasion in complex games is commonplace. While interactions between businesses are often modelled via Markov Decision Processes, in reality the game tree isn't surveyable because a player may make an offer that his opponent didn't foresee as a possible move. Similarly, in board games like *Civilisation* and *The Settlers of Catan* (or *Settlers*) there are unbounded options for trading due to, for instance, the capacity to promise a specific future trade: e.g., *I'll give clay for wheat now and ore when I get it if you don't block me*. So standard algorithms for computing preferences over the outcomes of the current negotiation, like backwards induction and its variants (Shoham and Leyton-Brown, 2009), break down (Cadilhac et al., 2013).

We therefore need a general method for exploring the benefits and risks of persuasion in contexts that go beyond the ones modelled in standard negotiation games or argumentation theory. We supply a method here, using game simulations among computer agents whose symbolic strategies differ in transparent and controlled ways. We identify the following: when a persuasion move is likely to be successful (i.e., the recipient is persuaded); when successful persuasion results in a higher chance to win the overall game; and conversely when attempts to persuade are ineffective in improving win rates, even if they're successful.

Our empirical set up provides a proof by demonstration that one can rapidly design, test and adapt symbolic persuasion strategies, with adaptation being guided by the quantitative performance metrics from game simulations. Specifically, we use our method to modify an existing agent that plays *Settlers*, and the result is a more effective player. Previous work on automatically learning *Settlers* strategies has shown that a decent prior player is critical for learning to succeed (Szita et al., 2010; Pfeiffer, 2003). We provide a principled way to build such priors, but investigating whether they enhance machine learning is future work.

In section 2 we describe related research, including work on agents that play *Settlers*. In section 3 we describe the rules of *Settlers* and the implemented agents that we use as a starting point. We then present our experiments, in which we manipulate the context in which the persuading agent chooses to perform a persuading move, the type of persuading move he attempts, and the strategies opponents adopt for accept or rejecting persuading moves. We provide quantitative metrics via game simulations of the effects of their different policies—e.g., their win rates and the number of persuasion moves executed. Our experiments radically discriminate among the persuasion strategies, identifying the strong strategies from the weak ones even though our game lacks any analytic solution.

2 Related Work

Negotiation in game theory (e.g., Binmore (1998), Brams (2003)) models when and how one suffers from the 'winner's curse' (i.e., overpaying for an item, given the opponents' preferences) and problems analogous to the prisoner's dilemma (i.e., can one player trust the other to voluntarily cooperate during negotiation). But since each player has a complete and static model of his own preferences over the outcomes of negotiation the scope of persuasion gets restricted to persuading an opponent to change his beliefs but not his preferences over trades. Consequently game-theoretic models of persuasion (e.g., Rubinstein and Glazer (2006)) focus on the problem of predicting the *credibility* of the persuasive move. We address different questions: if one isn't certain about which trades will help you, or hurt you, for winning the overall (complex) game, then how can one balance the benefits and risks of successfully persuading? And hence at what stage in a complex game is successful persuasion most likely to increase one's chances of winning the overall game? We propose an empirical method for answering these questions.

Our domain of study is *Settlers* (see section 3 for motivation). Empirical approaches to modelling *Settlers* deploy Monte Carlo Tree Search (Szita et al., 2010; Roelofs, 2012) and reinforcement learning (Pfeiffer, 2003). But even though they all use a simplified game, with no trading or negotiating, they all need a decent prior model for learning to succeed. So their priors encode sophisticated strategies, defined via complex *hand-coded heuristics*. Our work contributes to the general problem of developing decent priors: we supply an empirical framework where hand-coded heuristics can be rapidly designed and improved in light of quantitative performance metrics; e.g., Guhe and Lascarides (2013; 2014b) where we (a) identify negotiation strategies in *Settlers* that compensate for deficiencies in belief, e.g., memory loss, and (b) improve the building strategy used by our agents. Here, we identify effective persuasion strategies.

In trade negotiations, the persuading agent aims for either:

1. **More Trades:** i.e., a desired trade he might not achieve otherwise (e.g., *If you accept this trade, you'll get clay and be able to build a road*); or

2. **Fewer Opponent Trades:** i.e., he stops two opponents from trading with each other (e.g., *Don't trade with him! He's about to win!*)

Kraus and Lehmann (1995) propose hand-built symbolic strategies for performing both these kinds of persuasion moves within the complex game *Diplomacy*, but the individual aspects of the strategies aren't evaluated in controlled and transparent ways. We supply an empirical framework for doing just that. Here, we focus on game simulations for testing only those persuasion strategies that aim for More Trades; we address persuasion strategies aiming for Fewer Opponent Trades (FOT) in Guhe and Lascarides (2014a).

Achieving a successful persuasion move—i.e., one where the opponent is persuaded—is dependent on the persuading agent's ability to adapt his persuasive argument to the current context and his

type of opponent. In a game of imperfect information, some executed persuasion moves are unsuccessful; i.e., they fail to persuade. So in this paper we explore how the persuading agent's ability—or inability—to articulate arguments to opponents of various types should impact on his decisions about when to execute a persuading move so as to maximise his chances of winning the overall game.

3 *The Settlers of Catan* and *JSettlers*

The domain for our experiments is the board game *The Settlers of Catan* (or *Settlers*, (Teuber, 1995); `www.catan.com`). We chose it for its complexity: it is multi-player, partially observable, non-deterministic and dynamic; and further, with negotiations being conducted in natural language, the game's options are unbounded (see earlier discussion). Thus our experiments prove that one can rapidly design and improve persuasion strategies in a principled and empirically grounded way even when game-theoretic algorithms for optimisation break down.

Settlers is a win–lose board game for 2 to 4 players. Each player acquires resources (ore, wood, wheat, clay, sheep) and uses them to build roads, settlements and cities on a board shown in Figure 1. This earns Victory Points (VPs); the first player with 10 VPs wins. Players can acquire resources via the dice roll that starts each turn and through trading with other players—so they negotiate trades. Players can also lose resources: e.g., a player who rolls a 7 can rob from another player. What's robbed is hidden, so players are uncertain about their opponents' resources. Deciding what resources to trade depends on what you want to build; e.g., a road requires 1 clay and 1 wood. Because *Settlers* is a game of imperfect information, agents frequently engage in 'futile' negotiations that result in no trade; i.e., they miscalculate the equilibria (Afantenos et al., 2012).

Our experiments modify an existing *Settlers* playing environment and automated *Settlers* player called *JSettlers* (`jsettlers2`, Thomas (2003)). *JSettlers* is a client–server system: a server maintains the game state and passes messages between the players' clients, which can run on different computers. Clients can be humans or computer agents. Here, we report on simulations between computer agents.

The *JSettlers* agent goes through multiple phases after the dice roll that starts his turn:

Figure 1: A game of *Settlers* in *JSettlers*.

1. Deal with game events: e.g. placing the robber; acquiring or discarding resources.
2. Determine legal and potential places to build.
3. Find the *Best Build Plan* (BBP), viz. the agent's estimate of which build action gets him to 10 VPs in the shortest estimated time.
4. Try to execute the BBP, including negotiating and trading with other players.

Since we wish to study persuasion, our agents vary only in their policies for step 4, cf. section 4. Thomas (2003) describes steps 1–3. Here it only matters that the existing decisions on when to trade mean trading correlates with winning (Guhe and Lascarides, 2013).

In step 4 all agents have three existing possible responses to a trade offer: accept, reject or counteroffer. We equip our persuading agent with one more: to persuade an opponent to accept his trade offer. In our experiments, we vary the strategy for choosing among this expanded set of actions, and the strategies for reacting to the new option.

4 Evaluating Persuasion Moves

4.1 Motivation

There are a whole host of persuasive arguments that can accompany a trade offer—*Settlers* doesn't restrict the types of trades nor the reasons for trading in any significant way. A small selection of possible persuasion moves is:

(1) Give me 1 ore for 1 wheat and you can immediately build a settlement, which you can't build without the wheat.

(2) Give me 1 ore for 1 wheat and only then will you have enough wheat to make a trade with your 3:1 port.

(3) If you give me 1 ore for 1 wheat, you can use the wheat to trade for James' clay, so that you can build your road.

(4) If you give me 1 ore for 1 wheat, I won't rob you the next time I'm playing a knight card.

(5) If you give me 1 ore for 1 wheat, I'll build a road that blocks Nick from that port.

So the benefits and risks of persuasion will depend on (at least):

$\mathcal{P}$'s ingenuity: the range of contexts where the persuading agent (who we'll label $\mathcal{P}$) can articulate a persuasive move like those in (1) to (5) and beyond.

$\mathcal{P}$'s caution: In those contexts where his ingenuity provides a candidate persuasion move, $\mathcal{P}$'s strategy for deciding whether to actually make that move; and

$\mathcal{G}$'s gullibility: how inclined the recipient (labelled $\mathcal{G}$) is to accept $\mathcal{P}$'s persuasion move and hence also the trade offer.

Ingenuity and caution are distinct factors that determine a persuader's player type: ingenuity affects the persuader's range of options (he is more or less able to generate a candidate persuasion move); caution affects the persuader's penchant for actually executing a persuasion move when such a move is an option. Our experiments vary both factors, because the optimal level of caution may be different for an ingenious vs. non-ingenious agent—after all, an ingenious cautious player's behaviour is not in general equivalent to that of a non-ingenious, non-cautious player.

Asher and Lascarides (2013) show that a rational $\mathcal{G}$ will normally accept $\mathcal{P}$'s speech act—and a persuading move in particular—if $\mathcal{G}$ believes $\mathcal{P}$ to be *sincere* (i.e., $\mathcal{P}$ believes what he says) and *competent* (i.e., what $\mathcal{P}$ believes is true). But $\mathcal{P}$ can appear sincere and competent without actually being so. For instance, $\mathcal{P}$ can utter (1) but be ignorant about whether $\mathcal{G}$ has the other resources he needs for a settlement (i.e., clay, wood and sheep) and/or he may lack evidence that building a settlement is better for $\mathcal{G}$ than $\mathcal{G}$'s current build plan (whatever that is). In this case, $\mathcal{P}$ is neither sincere nor competent. But even if $\mathcal{G}$ lacks clay, wood and sheep, it's still consistent for him to assume that $\mathcal{P}$ was sincere (but inconsistent to assume he's competent), for $\mathcal{G}$'s resources aren't observable to $\mathcal{P}$ and $\mathcal{P}$'s beliefs aren't observable to $\mathcal{G}$. Further, if $\mathcal{G}$ does have clay, wood and sheep, then because $\mathcal{G}$ is uncertain about his own relative preferences over build plans, it's consistent for $\mathcal{G}$ to assume that $\mathcal{P}$ is both sincere and competent in (1)'s implicated content, that building a settlement is both possible *and* better for $\mathcal{G}$. Thus there's scope for $\mathcal{P}$ to successfully bluff, getting $\mathcal{G}$ to accept his persuasion move even though he's neither sincere nor competent. Our experiments thus investigate when bluffing succeeds, and whether successfully bluffing helps $\mathcal{P}$ win the overall game.

4.2 The Agents' Contexts

We start with a persuading agent $\mathcal{P}$ with *maximal ingenuity*—i.e., he can make a persuasion move every time he makes a trade offer and is unrestricted in the number of such moves he can make in the course of the game. Further, we make $\mathcal{G}$ *maximally gullible*: he assumes $\mathcal{P}$'s persuasion move is convincing so long as the proposed trade is executable. We then vary $\mathcal{P}$'s *caution*, by making $\mathcal{P}$ start executing persuasion moves only once the first agent reaches a specified number of VPs. We call this factor *VP*. In Guhe and Lascarides (2014a) we showed that the timing of persuasion moves is crucial and moves early and late in the game are much less effective than if they are used when the first player has reached around 7 VPs.

We call these agents *simple*. In terms of Guhe and Lascarides (2013) these agents are both *ignorant*, in that they use only observable information (VPs for $\mathcal{P}$, his own resources for $\mathcal{G}$) to decide what to do. A simple $\mathcal{P}$ is also relatively incautious, because the leader's VPs is the only factor that prevents $\mathcal{P}$'s trade offer from having an accompanying persuasion move too.

From this starting point, we will then vary $\mathcal{P}$'s degree of *caution*, by restricting the contexts (over and above *VP*) in which $\mathcal{P}$ actually chooses to make a persuasion move, and $\mathcal{G}$'s *gullibility* by restricting the contexts in which $\mathcal{G}$ accepts $\mathcal{P}$'s persuasion moves.

4.3 Method for Simulation and Analysis

A simulation for testing the different persuasion moves consists of 1 persuading agent ($\mathcal{P}$) playing 3 non-persuading opponents ($\mathcal{G}$) 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 run on the same machine, and 10,000 games take about 1 hour on a current desktop computer.

Apart from the agents' win rates, we measure how many persuasion moves $\mathcal{P}$ actually makes: the fewer persuasion moves $\mathcal{P}$ needs to gain a significant advantage in winning, the more efficient they are in achieving desirable effects.

We performed *Z*-tests with $p < 0.01$ to test significance of win rates against the null hypothesis. This means that win rates between 0.24 and 0.26 don't differ significantly from the null-hypothesis; so we highlight the 0.26 threshold in the graphs below. We report the average numbers for $\mathcal{P}$ for each simulation, and averages across all three of $\mathcal{P}$'s opponents. Due to the large number of games per simulation even small differences can be significant. At the same time, there were no significant differences between the three $\mathcal{G}$s. Persuasion does not affect the average length of the game, which is consistently between 21 and 21.5 rounds.

5 Simple vs. Cautious $\mathcal{P}$

5.1 $\mathcal{P}$:$\varnothing$ vs. $\mathcal{P}$:*PB*

In the first set of simulations we compared simple agents (i.e. agents the make/accept the maximum number of persuasion moves) and then restricted $\mathcal{P}$ to a more self-serving context:

1. ***None* ($\varnothing$):** $\mathcal{P}$ using this context makes a persuasion move with every trade offer *proviso* the *VP* factor; $\mathcal{G}$ using this context accepts all persuasion moves and the accompanying trade offer if the trade is executable (i.e. $\mathcal{G}$ has the resources for making the trade).
2. ***Persuader Build* (PB):** $\mathcal{P}$ makes the persuasion move iff *VP* is satisfied *and* the proposed trade allows him to build immediately, i.e. to execute his BBP after making the trade.

A $\mathcal{P}$ who adopts *PB* is relatively cautious: he's attempting to mitigate the risk of his deficient preferences over trades by ensuring that all successful persuasions result not only in his desired trade but also in the immediate benefit of building.

Figure 2 shows the simulation results for the configuration ($\mathcal{P}$:$\varnothing$, $\mathcal{G}$:$\varnothing$)—i.e. $\mathcal{P}$ can make an unlimited number of persuasion moves ($N = \infty$) and $\mathcal{G}$ accepts all such moves—as well as for the configuration ($\mathcal{P}$:*PB*, $\mathcal{G}$:$\varnothing$). $\mathcal{P}$:*PB*'s win rates are almost as good as $\mathcal{P}$:$\varnothing$'s (0.363 vs. 0.377 at 2 VPs; 0.274 vs. 0.285 at 8 VPs) but he needs substantially fewer persuasion moves for this (15.4 vs. 40.8 at 2 VPs; 1.4 vs. 6.0 at 8 VPs).

Realistically, a fully ingenious $\mathcal{P}$ risks irritating his opponents and making them suspicious if he makes a persuasion move every time he can—even 15 moves in the course of a game (cf. $\mathcal{P}$:{*PB*, *VP* = 2}) is more than humans do according to our corpus data (Afantenos et al., 2012). Figure 3 shows what happens if the number of persuasion moves $\mathcal{P}$ can make are limited ($\mathcal{P}$:{$N \in [1, 3]$}, $\mathcal{G}$:$\varnothing$). $\mathcal{P}$:*PB* achieves a significant improvement over the null-hypothesis even when he only makes 1 move at most, so long as he makes that move after the first player reaches 6 VPs. (This is consistent with our results in Guhe and Lascarides (2014a).) The less cautious $\mathcal{P}$:$\varnothing$ needs to be able to make at least 3 moves to gain a significant advantage. The right graph in Figure 3—depicting the number of moves $\mathcal{P}$ actually made—also shows that even though $\mathcal{P}$:*PB* makes fewer moves than $\mathcal{P}$:$\varnothing$, he achieves a much higher win rate. So perhaps surprisingly, the less ingenious $\mathcal{P}$ needs to be more cautious.

In the following, we will only report on simulations where $\mathcal{P}$ can make an unlimited number of persuasion moves (i.e., $N = \infty$), because the main effect for N is the same across simulations: the higher N is, the more moves $\mathcal{P}$ makes and the more games he wins.

5.2 Number of gullible agents

An agent's success is always highly dependent on his opponents. So we checked how much $\mathcal{P}$'s performance depends on the number of persuadable opponents he plays against. These simulations vary the number of $\mathcal{G}$ opponents who accept persuasion moves vs. those (non-$\mathcal{G}$) opponents who never accept them. For conditions ($\mathcal{P}$:$\varnothing$, $\mathcal{G}$:$\varnothing$) and for ($\mathcal{P}$:*PB*, $\mathcal{G}$:$\varnothing$), $\mathcal{P}$ retained a big advantage over all three opponents even when only one of them is persuadable. Further, deploying *PB* helps $\mathcal{P}$ achieve almost the same win rate as without it, but with fewer than half of the persuasion moves.

	wins		moves made	
config.	$\varnothing$	*PB*	$\varnothing$	*PB*
3 $\mathcal{G}$	0.383	0.363	40.7	15.4
2 $\mathcal{G}$	0.342	0.341	47.6	19.1
1 $\mathcal{G}$	0.302	0.315	60.3	24.7

The reason why the persuader needs more moves the fewer opponents are gullible is that

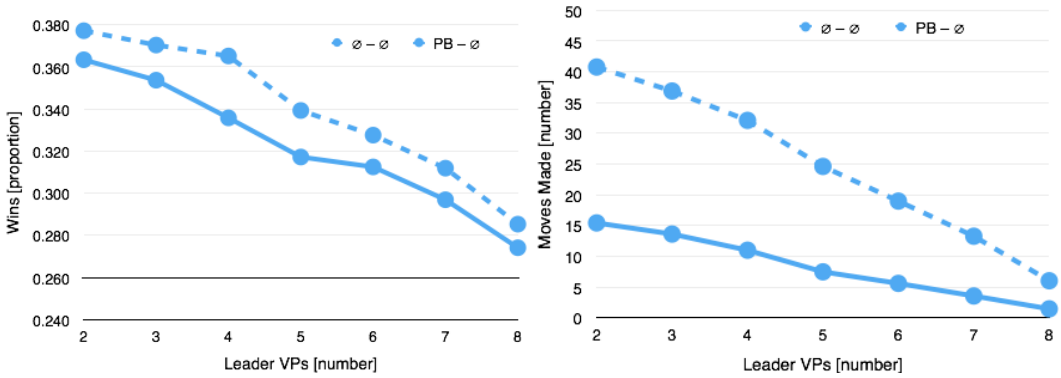


Figure 2: Win rate and persuasion actually moves made, against the *VP* factor (i.e., the leader's minimum VPs before persuading can start). The dashed line is $\mathcal{P}$:$\varnothing$, the solid line is $\mathcal{P}$:*PB*.

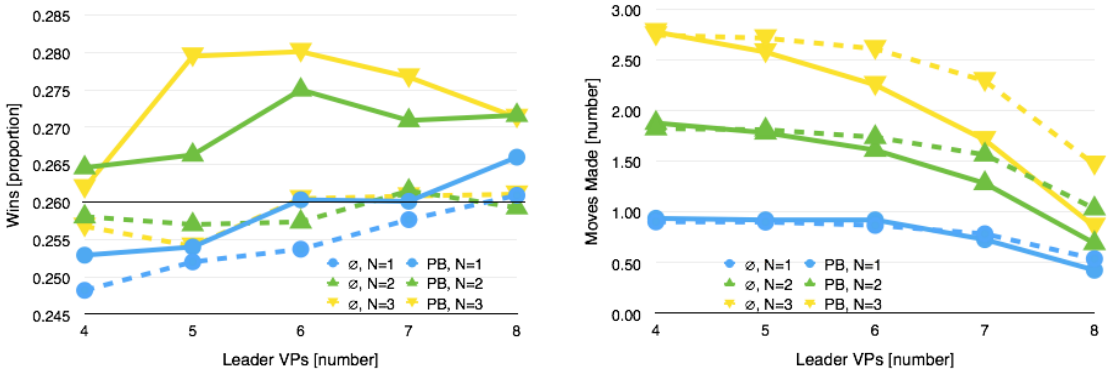


Figure 3: Win rate and moves made against the *VP* factor. Dashed lines are $\mathcal{P}$:$\varnothing$ and solid lines $\mathcal{P}$:*PB*.

more of the persuader's trade offers are unsuccessful (non-gullible agents accept offers at a normal rate). So $\mathcal{P}$ has to make more offers to get the trades he wants.

6 A more discerning $\mathcal{G}$

So far, our $\mathcal{G}$ agents are so gullible that they don't test the persuasive argument for sincerity or competence. We now restrict $\mathcal{G}$'s gullibility: instead of accepting all persuasion moves where the trade is executable ($\mathcal{G}$:$\varnothing$), we make $\mathcal{G}$ accept whatever $\mathcal{P}$'s persuasive move is only if $\mathcal{G}$ can build something or make a bank/port trade as a result of trading (in the following we abbreviate *trade with the bank or an available port* to *bank trade*). In other words, factors for $\mathcal{G}$ accepting a persuasion move are:

1. ***Gullible Build* (GB):** $\mathcal{G}$ accepts the persuasion move only if it enables him to build a type of piece that he cannot build without making the trade.
2. ***Gullible Bank Trade* (GBT):** $\mathcal{G}$ accepts the persuasion move only if after making the trade he can make a bank trade immediately.
3. **GBoBT:** The disjunction of these two cases.

Note that $\mathcal{G}$:*GB* by default assumes that $\mathcal{P}$ is sincere and competent on persuasive moves like (1), and $\mathcal{G}$:*GBT* by default assumes that $\mathcal{P}$ is sincere and competent on persuasive moves like (2).

Here it is important to distinguish the *persuasion move* from the *trade offer* that it is accompanying: Even if $\mathcal{G}$ does not accept the persuasion argument (e.g., $\mathcal{G}$ infers $\mathcal{P}$'s persuasion argument is not competent), he will still evaluate the trade offer in it's own right. For example, in (1), $\mathcal{G}$ may still agree to exchange 1 ore for 1 wheat, even if this does not enable him to immediately build the settlement as $\mathcal{P}$ claims. That is, $\mathcal{G}$ never rejects a trade offer with a persuasion move if he would have accepted it without the persuasion.

Figure 4 gives the results for both $\mathcal{P}$:$\varnothing$ and $\mathcal{P}$:*PB*. In all cases, $\mathcal{P}$ fares better in the *PB* context than in the $\varnothing$ one and with fewer persuasion moves; i.e., $\mathcal{P}$:*PB* is not only more effective but

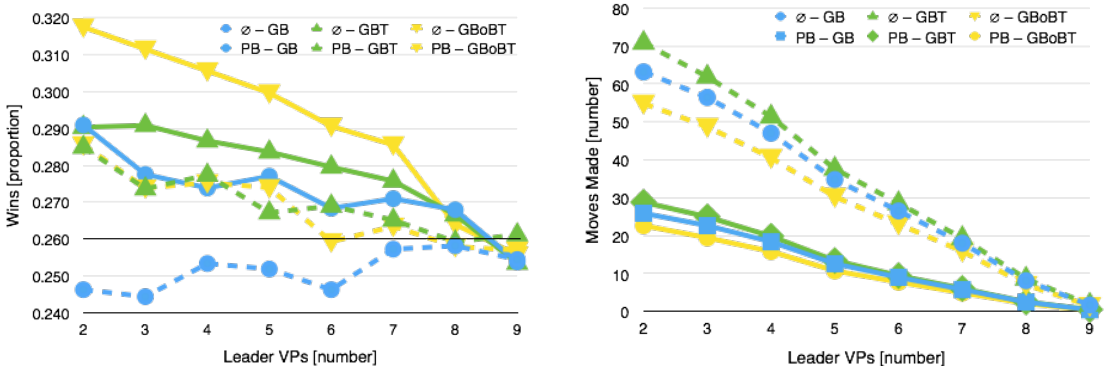


Figure 4: Win rate and moves made when varying $\mathcal{G}$'s gullibility. Dashed lines are $\mathcal{P}$:$\varnothing$; solid lines $\mathcal{P}$:*PB*.

also more efficient. Indeed, there is no effect for ($\mathcal{P}$:$\varnothing$, $\mathcal{G}$:*GB*)—i.e., no agent gains an advantage—but there is an effect for ($\mathcal{P}$:*PB*, $\mathcal{G}$:*GB*). So as $\mathcal{G}$ gets less gullible, $\mathcal{P}$ should get more cautious (i.e., play with strategy *PB*). On the other hand, as we saw in sections 5.1 and 5.2, so long as at least one of $\mathcal{P}$'s opponents is *maximally gullible* ($\mathcal{G}$:$\varnothing$), $\mathcal{P}$ should be maximally incautious (i.e., $\mathcal{P}$:$\varnothing$).

For $\mathcal{G}$:*GBT*, $\mathcal{P}$ gains an advantage for both of his contexts. Thus, the potential benefit for $\mathcal{G}$ if $\mathcal{P}$ makes move (2) is smaller than that for move (1); conversely, $\mathcal{P}$'s risk in successfully making move (1) is smaller than that for move (2). When $\mathcal{G}$ accepts persuasion attempts of both kinds so long as it's consistent with sincerity and competence—i.e. $\mathcal{G}$:*GBoBT*—then $\mathcal{P}$ has a bigger advantage than if $\mathcal{G}$ uses only one of the contexts, and $\mathcal{P}$ needs even fewer moves. So, one general observation here is that the more types of persuasion moves $\mathcal{G}$ accepts, the more successful $\mathcal{P}$ is and the fewer moves $\mathcal{P}$ needs to achieve this.

7 $\mathcal{P}$ taking $\mathcal{G}$'s Context into Account

So far, $\mathcal{P}$ does not reason about $\mathcal{G}$'s likely reaction when deciding whether to make a persuasion move. But as we said earlier, persuasion must appear sincere and competent to a rational $\mathcal{G}$ to be successful. And $\mathcal{P}$ can reduce the risk of miscalculating equilibria and making futile moves by reasoning about $\mathcal{G}$'s likely reaction. We investigate this by restricting $\mathcal{P}$'s ingenuity—he can only articulate moves of the form (1) or (2)—and $\mathcal{P}$'s caution in the following ways:

1. ***Persuader Opponent Build*** **(POB):** $\mathcal{P}$ only makes a persuasion move only if he believes that it allows $\mathcal{G}$ to build something that he cannot build without making the trade.
2. ***Persuader Opponent Bank Trade*** **(POBT):** $\mathcal{P}$ makes the persuasion move only if $\mathcal{P}$ believes that after making the trade, $\mathcal{G}$ can immediately make a bank trade that he cannot make without the trade.
3. **POBoOBT:** The disjunction of these cases.

Whether $\mathcal{P}$ executes a persuasion move now depends on $\mathcal{P}$'s beliefs about $\mathcal{G}$'s resources. For instance, agent $\mathcal{P}$:*POB* must believe that the resources $\mathcal{G}$ gets in the proposed trade are necessary and sufficient for $\mathcal{G}$ to immediately build. So $\mathcal{P}$:*POB* is in effect only making persuasion moves of form (1), and executes such a move only if $\mathcal{P}$ believes that a $\mathcal{G}$ player of the following type will accept it: (a) $\mathcal{G}$ is rational, and so accepts a move iff $\mathcal{G}$ believes it's sincere and competent; and (b) $\mathcal{G}$ defaults to believing moves are sincere and competent. Similarly, $\mathcal{P}$:*POBT* only makes persuasion moves of the form (2) and only executes them if $\mathcal{P}$ believes a $\mathcal{G}$ player of the above type will accept it; $\mathcal{P}$:*POBoOBT* is slightly more ingenious, using persuasion moves of both types.

The agents use the standard *JSettlers* belief model, i.e. no memory loss and fully accurate beliefs about how many resources each opponent has, but some are of unknown type because of robbing. In terms of Guhe and Lascarides (2013), $\mathcal{P}$ is relatively cautious: he does not take $\mathcal{G}$'s unknown resources into account, i.e. he only makes a persuasion move, if he *knows* that $\mathcal{G}$ can execute the build or bank trade he promises—$\mathcal{P}$ does not bluff.

Depending on his gullibility configuration, $\mathcal{G}$ accepts different persuasion arguments, e.g. $\mathcal{G}$:*GB* is only susceptible to the arguments of $\mathcal{P}$:*POB* (or, $\mathcal{P}$:*POBoOBT*) but not $\mathcal{P}$:*POBT*.

Similar to the previous result, $\mathcal{P}$:*POB* does not improve his win rate but $\mathcal{P}$:{*PB*, *POB*} does. And

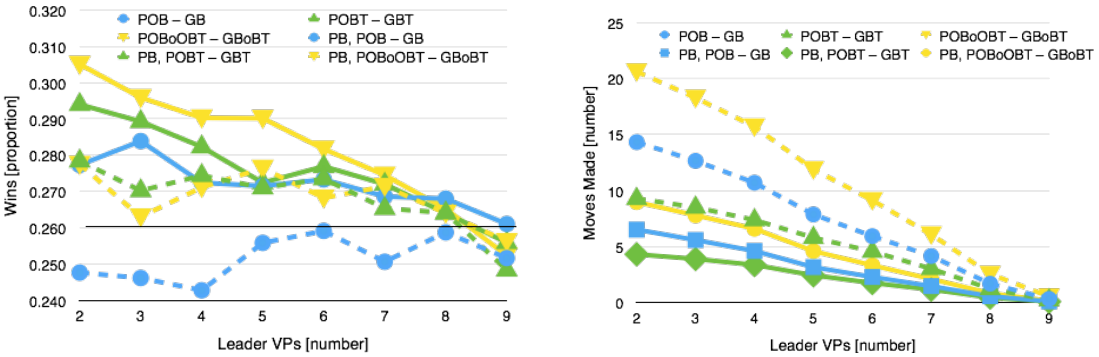


Figure 5: Win rate and moves made when $\mathcal{P}$ takes $\mathcal{G}$'s context into account. Dashed lines are $\mathcal{P}$:$\varnothing$; solid lines $\mathcal{P}$:*PB*.

when being more selective about making persuasion moves (by adopting *PB*), adding *POB* does not reduce $\mathcal{P}$'s win rate, but he needs only about a quarter of the moves.

In the ($\mathcal{P}$:{*PB*, *POBT*}, $\mathcal{G}$:*GBT*) context, $\mathcal{P}$'s win rate is similar to the one without $\mathcal{P}$ taking $\mathcal{G}$'s context into account. This strategy is more efficient for $\mathcal{P}$ (fewer moves) and more effective (higher win rate) than *POB*. Again, the *PB* context is more effective and efficient than the $\varnothing$ context.

Finally, in ($\mathcal{P}$:{*PB*, *POBoOBT*}, $\mathcal{G}$:*GBoBT*) $\mathcal{P}$ makes both kinds of persuasive moves as well as both kinds of assessments about $\mathcal{G}$'s state, and $\mathcal{G}$ is selective about both types of moves. The added opportunities that $\mathcal{P}$ obtains through his increased ingenuity compared to an agent who can make only one type of argument leads to more persuasive moves being executed and a higher win rate.

Comparing these results to the simulations when $\mathcal{P}$ does not take $\mathcal{G}$'s gullibility into account, we again see that $\mathcal{P}$ can increase its efficiency (he makes fewer moves) without sacrificing his effectiveness (the win rates do not differ substantially).

8 Conclusions

In this paper we used *The Settlers of Catan* to investigate the power of persuasion in a multi-player, partially observable, non-deterministic, dynamic, unbounded game. We established an empirical method involving game simulations, with the heuristics that the persuading agent and his recipients use being evaluated in controlled ways and improved upon.

We found that the more ingenuity the persuader has at articulating persuasive arguments, and the more gullible his recipients, the more successful he becomes at winning the overall game. Indeed, one gullible agent is sufficient for the persuader to gain an advantage over all three opponents. If he lacks ingenuity and so is restricted to only certain kinds of arguments, then it helps to make performing a persuasive move dependent on whether the proposed trade will enable him to immediately build. The persuader can also increase the proportion of his persuasion moves that are successful without harming his win rate by reasoning about how his opponent will react.

Gullible agents, who assume the persuader is sincere and competent by default, cannot improve over the null-hypothesis—a 25% win-rate. But they 'lose less' if they are selective about the persuasion moves they comply with; here, if you comply with just one kind of persuasion move, it should be the one like (1) (i.e., you can immediate build but only if you execute the proposed trade).

We are currently collecting data on how persuasive human opponents find the More Trade persuasion moves we investigated here and will then investigate persuasion that aims for Fewer Opponent Trades. We will then use our *Settlers* environment to test our best persuasive agents against humans. We will also investigate the impact of our improved priors on automatically learning *Settlers* strategies and opponent modelling similar to the work by Gal et al. (2004) in order to adapt to human opponents over the course of a game.

Acknowledgments

We would like to thank the reviewers for helpful comments. This work is funded by ERC grant 269427 (STAC).

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cedric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Vladimir Popescu, Verena Rieser, and Laure Vieu. 2012. Modelling strategic conversation: model, annotation design and corpus. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (Seinedial)*, Paris.

Leila Amgoud and Srdjan Vesic. 2014. Rich preference-based argumentation frameworks. *International Journal of Approximate Reasoning*, 55.

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Sematics & Pragmatics*, 6(2):1–62.

Ken Binmore. 1998. *Game Theory and the Social Contract: Just Playing*. MIT Press.

Steven Brams. 2003. *Negotiation Games: Applying game theory to bargaining and arbitration*. Routledge.

Anaïs Cadilhac, Nicholas Asher, Alex Lascarides, and Farah Benamara. 2013. Preference change. submitted.

Ya'akov Gal, Avi Pfeffer, Francesca Marzo, and Barabara J Grosz. 2004. Learning social preferences in games. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI), San Jose, California, July 2004*.

H. Paul Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press.

Markus Guhe and Alex Lascarides. 2013. Effects of belief and memory on strategic negotiation. In *Proceedings of SEMDIAL 2013 – The 17th Workshop on the Semantics and Pragmatics of Dialogue, Amsterdam, 16–18 December 2013*, pages 82–91.

Markus Guhe and Alex Lascarides. 2014a. The effectiveness of persuasion in The Settlers of Catan. In *Proceedings of CIG 2014, The IEEE Conference on Computational Intelligence and Games*.

Markus Guhe and Alex Lascarides. 2014b. Game strategies for The Settlers of Catan. In *Proceedings of CIG 2014, The IEEE Conference on Computational Intelligence and Games*.

Sarit Kraus and Daniel Lehmann. 1995. Designing and building a negotiating automated agent. *Computational Intelligence*, 11(1):132–171.

Michael Pfeiffer. 2003. Machine learning applications in computer games. Master's thesis, Technical University of Graz.

Gijs-Jan Roelofs. 2012. Monte Carlo Tree Search in a modern board game framework. Research paper available at `umimaas.nl`.

Ariel Rubinstein and Jacob Glazer. 2006. A study in the pragmatics of persuasion: a game theoretical approach. *Theoretical Economics*, 1(4):395–410.

Ariel Rubinstein. 2007. What is the proper role of game theory to other disciplines? In Vincent F Hendricks and Pelle Guldborg Hansen, editors, *Game Theory: 5 Questions*. Automatic Press.

Yoav Shoham and Kevin Leyton-Brown. 2009. *Multiagent Systems: Algorithmic, Game-Theoretic and Logical Foundations*. Cambridge University Press.

István Szita, Guillaume Chaslot, and Pieter Spronck. 2010. 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.

Klaus Teuber. 1995. *Die Siedler von Catan: Regelheft*. Kosmos Verlag, Stuttgart, Germany.

Robert S Thomas. 2003. *Real-time Decision Making for Adversarial Environments Using a Plan-based Heuristic*. Ph.D. thesis, Northwestern University.

Signaling Non-Speaker commitment in Transparent Free Relatives: A paired Speaker-Hearer judgment study.

Jesse A. Harris
University of California, Los Angeles
jharris@humnet.ucla.edu

Abstract

In a typical conversation, Speakers are assumed to be committed to the content of their utterances. Recent research has uncovered several linguistic expressions or prosodic contours that convey subtle interactions between the commitments of discourse agents and the presumed source of the information. Another such case is that of Transparent Free Relatives, as in *That's an instance of what pragmaticians call 'implicature'*, which provide a systematic way to explicitly introduce a source (*pragmaticians*) into an attribution statement (*call 'implicature'*), but may also leave the source unexpressed, as in *That's an instance of what is called 'implicature'*. We explore the factors that give rise to Non-Speaker commitment in a novel two-person experimental paradigm, finding that (i) the presence of an explicit third person source and (ii) the tense of the attribution statement provide reliable cues to Non-Speaker commitment.

1 Introduction

The terms that a Speaker uses to describe an object or event often convey an implicit point of view, the connotations of which presumptively default to Speaker commitment or acceptance, unless there are clear cues to the contrary (Grice, 1978; Levinson, 2000; Harris and Potts, 2009). Speakers may selectively maneuver this default by modifying the *means* by which a potentially controversial element is designated. Here, the term *beergarita* (a literal and linguistic blend of *beer* and *margarita*) is enveloped in a so-called Transparent Free Relative (1b), which raises the issue of whether the Speaker believes the term *beergarita* is appropriate, relative to the canonical variant (1a).

(1) a. John made Mary a beergarita.

b. John made Mary *what he calls a beergarita*.

Syntactically, Transparent Free Relative (TFR) clauses are Free Relative (FR) clauses that 'transparently stand in' for some constituent contained within the FR clause itself (Wilder, 1998; Grosu, 2003). In example (1b), the phrase *what he calls a beergarita* is the TFR clause, and the underlined noun *beergarita* is the element for which the entire clause substitutes, which often has a quotational or indirect speech effect. Pragmatically, TFRs introduce a term or expression (*beergarita*) through an explicit source (*John*) for the attribution (*he calls*). TFRs thus provide a systematic way to modulate the degree to which a Speaker conveys her commitment to a term.

After reviewing current research on commitment in pragmatics, we turn to a brief overview of the pragmatics of TFRs, followed by a description of a novel two-person experiment that tests the basic predictions of a cue-based account, in which multiple interacting cues work together to promote an interpretive shift from away from Speaker commitment (Smith, 2003; Harris and Potts, 2009).

2 Speaker commitment

Speakers presumably believe what they say, or at least intend to convey as much, unless their utterances are otherwise marked. In other words, they are assumed to be *committed* to the content of their utterance (Hamblin, 1971; Levi, 1991). We use the term 'commitment', in the sense that a discourse agent α may be committed to the (propositional) content ϕ expressed by an expression E when α makes public, in some way, α's belief in ϕ through the use of E. A commitment differs from a *genuine belief* in that (i) commitments are necessarily public, and hence can be expected to generate implicature of the usual sort, and to li-

cense discourse moves, and (ii) commitments may be disingenuous, in that one may adopt a commitment for, say, the sake of polite conversation or deception, among other reasons (Hamblin, 1970).

Of course, discourse agents need not share the commitments of others in the discourse. Consequently, discourse agents – and models of discourse – need to somehow pair discourse agents with their commitments in order to draw appropriate inferences. This is unlikely to be an easy task. A great deal of varied information must go into assessing the commitments of our conversational partners. Presumably, discourse agents rely on linguistic conventions, coupled with general knowledge about the discourse and the agents therein, to form reasonable approximations of another agent's commitments.

Several recent studies have investigated how particular lexical items, syntactic configurations or intonational contours interact with the commitments of agents in the discourse. Examples include rising declaratives (Bartels, 1997; Gunlogson, 2001), discourse particles (Farkas and Bruce, 2010), predicates of personal taste (Lasersohn, 2005; Malamud and Stephenson, 2011), polarity rises (Malamud and Stephenson, 2011), and expressive terms (Potts, 2005; Harris and Potts, 2009). For example, Gunlogson (2008) observes that rising declaratives typically require the Addressee to be publicly committed to the proposition under question. In (2), B has no reason for thinking that A would be committed to the proposition that the fruit she is eating is a persimmon; hence, B's use of the rising declarative sounds infelicitous in the context. Once A makes that commitment public, a rising declarative addressing the Addressee's commitment is licensed (3).

(2) A. (Coworker silently eating a piece of fruit.)

B. # That's a persimmon?

(3) A. This is the best persimmon I've ever tasted.

B. That's a persimmon?

Following Hamblin, Gunlogson (2008) proposes that every discourse agent has a set of publicly available discourse commitments, which may be modeled as the set of worlds which conform to those beliefs:[1]

(4) $cs_{\alpha,d} = \{w \in W$: all discourse commitments of agent α in discourse d are true in $w\}$

The discourse context C, at a particular point in time, can be represented as a tuple of such commitment sets for all agents in the discourse: $C_d = \langle cs_{\alpha,d}, cs_{\beta,d}, \ldots\rangle$. The common ground – mutually held beliefs about the world that unfold throughout a discourse – is then to be understood as the intersection of individual commitment sets.

As Gunlogson and others realized, however, the more complicated case of *implicit* commitment presents itself. In example (5), whether or not the Speaker is committed to the identification is left vague or underdetermined by the semantics. Provided that John is a reliable source, (5) could be used to indicate that the proposition *that's a persimmon* is likely correct. For example, if John is an expert gardener, I'm surely going to trust his judgment by default. However, if John is contextually understood to be largely ignorant about such things, the intuition is that (5) becomes a comment on John's beliefs, from which the Speaker must now take pains to distance herself.

(5) According to John, that's a persimmon.

Additionally, John's reliability may simply not be known. The Speaker may use the *According to John* clause to identify her *source* of information, without necessarily commiting one way or the other. Discourse agents may require more information regarding John's reliability before accepting (or rejecting) the statement into the common ground (Farkas and Bruce, 2010; Malamud and Stephenson, 2011).

I will classify such cases as *Non-Speaker commitment* even though there are surely important distinctions to be explored further. The case in which John is ignorant about gardening might more accurately be called *Speaker Non-commitment*, in that the Source, not the Speaker, is committed to the attitude. The case in which John's reliability is unknown is vague with respect to Speaker commitment. Hence, Speaker commitment and Non-Speaker commitment need not be incompatible: a Non-Speaker source can serve as a proxy for the Speaker, as discussed below.

[1] We may assume for simplicity that $cs_{\alpha,d}$, and any update to it, results in a consistent model (Gunlogson, 2008). Similar constraints holds for standard models of common ground (Lewis, 1969; Fagin et al., 1995; Stalnaker, 2002). Possible worlds are used for convenience without commitment to their adequacy in capturing the finer points of belief or belief revision.

In a case similar to (5), Simons (2007), among many others, discusses the *evidential* use of embedding attitude predicates, such as *thinks*, *believes*, *imagines*, and so on.

(6) A. [Context: Pointing to a piece of fruit.] What is that?

B. i. That's a persimmon.

ii. I think/believe that's a persimmon.

iii. That, I think/believe, is a persimmon.

iv. John thinks/believes that's a persimmon.

The direct answer (6B.i) conveys a high degree of Speaker certainty, and thus complete Speaker commitment. First person embedding predicates (6B.ii–iii) function as hedges, allowing the Speaker to introduce some uncertainty regarding the accuracy of the statement. Finally, third person embedding cases defer the relevant attitude state to a Non-Speaker agent (*John*), triggering the inference that the Speaker is not in an appropriate epistemic state to provide an answer.

Such cases underscore the need to associate a commitment with a *source* for the content, defined by Gunlogson (2008) as follows:

(7) An agent α is a **source** for a proposition ϕ in a discourse d iff:

a. α is committed to ϕ; and

b. According to the discourse context, α's commitment to ϕ does not depend on another agent's testimony that ϕ in d.

Gunlogson proposes that commitments have sources. Sources may be the Speaker herself, or another discourse agent, such as the Addressee in the case of rising declaratives (2–3) or a third party mentioned in the sentence (5). In such cases, α's commitment might be said to be a *dependent commitment*:

(8) An agent α has a **dependent commitment** to a proposition ϕ in a discourse d iff:

a. α is committed to ϕ; and

b. According to the discourse context, α is not a source for ϕ in d.

Provided that an alternate source is not specified, a plausible interpretation takes the speaker to be the source of the claim, all else being equal. We may codify this intuition into the following presumptive inference:

(9) **Speaker commitment by default:** Unless otherwise indicated, assume that a Speaker is committed to content ϕ expressed in E.

This is a direct result of Grice's Maxim of Quality (roughly, "Do not say what you believe to be false or do not have evidence for"); in general, if speakers are expected to say what they have evidence for, then they should be likewise committed to the content of their reports.

We take it that discourse agents rely on *cues* from various sources to signal a contravention of default Speaker commitment (Smith, 2003; Harris and Potts, 2009), a position which raises a number of additional questions, including the following:

(10) i. What cues signal a Non-Speaker commitment to ϕ?

ii. How reliable are such cues?

iii. How do these cues interact? Do multiple cues work together to better signal Non-Speaker commitment? If so, are some cues stronger or more reliable than others?

We now turn to Transparent Free Relatives as a case study in this area in order to begin addressing these questions.

3 Transparent Free Relatives

Transparent Free Relatives (11b) are a type of Free Relative (11a) structure which serve to introduce a term or expression through predicates like verbs of saying, such as *call* or *describe as*, that select for equatives or small clauses, or else a clausal hedge, such as *appear to be* or *seem to be*. Like other types of FRs (Bresnan and Grimshaw, 1978; Caponigro, 2003), TFRs can stand in for many kinds of syntactic categories, but stand in most often for NPs. Although TFRs have a number of interesting syntactic and semantic properties (Wilder, 1998; Grosu, 2003; Schelfhout et al., 2004), those are not reviewed here.

(11) a. In the divorce hearing, John gave Mary [$_{\text{FR}}$ what she wanted].

b. In the divorce hearing, John gave Mary [$_{\text{TFR}}$ what she thinks of as reparations].

The examples below illustrate the most common use of TFRs, which are in abundance in news reporting, in which the commitments associated with the term shift to a Non-Speaker agent.

In (12), the use of the politically charged term *amnesty* is clearly ascribed to *Ted Cruz* in his description of the Democrats' proposal, contributing to a global perspective shift (Harris and Potts, 2009) in which evaluative terms like *right* reflect the point of view of Cruz, rather than of the reporter or the Senate Democrats. In (13), it is clear that the phrasing of the report reflects the attitude holder (Cummings), leaving the reporter's own commitments somewhat vague.

(12) Speaking Wednesday with conservative radio host Rush Limbaugh, Ted Cruz said that by promoting what he called "amnesty" for immigrants in the U.S. illegally, Senate Democrats are indeed hoping to get a lot more Democratic voters – but not among immigrants who did things the right way, like his father. (NPR: 20 May, 2013)

(13) But Cummings was not so happy about a media buildup to the hearing with what he called unfounded accusations aimed at smearing public officials. (NPR: May 09, 2013)

Pragmatically, however, TFRs are compatible with multiple interpretations besides a Non-Speaker perspective. Whether the Speaker accepts the appropriateness of the term *beergarita* depends, in part, on the extent to which John is deemed a trustworthy or authoritative source, and whether the Speaker is willing to adopt the term in question. Furthermore, authoritative sources can also be used to *introduce* the term to an ignorant audience, rather than to reject it; for example, *what we mixologists call a beergarita* identifies the Speaker as an authority, just as *what I would call a beergarita* can be understood as idiosyncratic or original to the Speaker. Additional factors such as modality, intonational marking, and non-verbal indicators such as head tilt or eyebrow raising may also play a role in establishing Non-Speaker commitment (Harris and Potts, 2011).

From among the many potential factors leading to Non-Speaker commitment, we concentrated on just two: (i) the presence of a third person source and (ii) the tense of the report, following previous findings that present tense promotes Non-Speaker interpretations of attitude reports in extended discourse contexts in comparison to past tense (Harris, 2012). In the case of TFRs, the present tense generates a habitual or generic interpretation of the attributive statement, suggesting that the attribution reflects a consistent commitment. In contrast, the past is consistent with an episodic reading, indicating that the attribution may not reflect a long-term belief, in addition to the habitual reading.

Although the variations in (14) are all ambiguous, they differ in whether we can attribute the term *beergarita* to a specific source (*John*) and whether the mode of reference is habitual (*calls*, *is called*) or possibly episodic (*called*).

(14) John made Mary what { is called / he called / he calls } a beergarita.

We predicted that the presence of a TFR would be insufficient, by itself, to overturn the Speaker default, but that the presence of a third person source would be a more important indicator. We also expected that the third person source would more greatly contribute to Non-Speaker interpretations when coupled with a present tense predicate, and that the combination of such cues would lead to more reliable interpretations between Speakers and Hearers.

4 Speaker-Hearer judgment task

This section introduces the results of a paired Speaker-Hearer experiment, in which two subjects participate in an interpretation judgment task.[2]

4.1 Materials and method

Fifteen pairs of subjects from UMass Amherst participated in the study (for a total of 30 subjects). Subjects were randomly assigned a role (Speaker or Hearer) prior to the experiment, and were seated facing away from one another, so that facial cues and gestures would not be a factor in the task.

Subjects were presented with 12 triplets of the form of (15), manipulating the presence of a Source (*Src*, *No-Src*) in the TFR and the Tense of the TFR predicate (*Present*, *Past*). The three conditions consisted of (i) No Source-Present (No-Src; *is called*), meant to establish a baseline for Speaker commitment with the construction, (ii) Source-Past (Src-Past; *he called*), giving one cue

[2]The terms 'Speaker' and 'Hearer' only indicate labels for the roles in the experiment. While it is expected that these roles would generalize, to a limited extent, to real conversation, it is also acknowledged that the 'Speaker' was reading the script, rather than articulating his or her own intention.

for Non-Speaker commitment, and (iii) Source-Present (Src-Pres; *he calls*), giving multiple cues for Non-Speaker commitment.

(15) John gave Mary what ...

a. **is called** a beergarita. *(No-Src)*

b. **he called** a beergarita. *(Src-Past)*

c. **he calls** a beergarita. *(Src-Pres)*

(16) *How did you interpret that sentence?*

i. Only John calls it a 'beergarita'. *(NSpO)*

ii. Everyone calls it a 'beergarita'. *(SpO)*

Items were presented in counterbalanced individually randomized order, so that subjects saw or heard one and only one item from each triplet, interspersed with 42 items from unrelated experiments (though all items asked about commitment in some form or another). Items were constructed so that only a quarter of the items contained potentially unfamiliar terms in the TFR, using a variety of attitude predicates: *call*, *think*, *believe*, *consider*, and *expect*. All items are provided in the appendix.

After Speakers read the item silently, they chose between two responses to an interpretation question like (16). As discussed above, Non-Speaker commitment is sometimes vague with respect to whether the Speaker would also endorse the attitude. The responses were constructed to be as unambiguous as possible, so that the Non-Speaker Oriented response (NSpO; 16.i) was phrased in terms of the stronger Speaker Non-commitment interpretation. The Speaker Oriented response (SpO; 16.ii) was intended cover all other interpretations, most prominent of which is Speaker commitment, by hypothesis. Order of responses was individually randomized for each participant.

After responding to the interpretation question, Speakers were asked to perform the item as though they were having a conversation, and their speech was recorded on a head-mounted microphone. The instructions to the Speaker included the following directions:

> You should think of this experiment as "a mind reading game" in which you report on what someone else has said. Your goal is to convey whether you also believe what you report on, while speaking as naturally as possible.

Hearers then made a judgment on the same interpretation question from the Speakers' performance alone – i.e., they responded to the question (16) without seeing additional text. The paradigm thus allows us to explore additional measures not typically gathered in similar experiments; in addition to interpretations and voice recordings, we also have a measure of Speaker-Hearer agreement, allowing us to determine precisely what factors reliably signal Non-Speaker commitment.

Items were presented with Linger (Rohde, 2003), which recorded responses from both Speaker and Hearer, as well as the audio performance of the Speaker. Each experimental session typically lasted no more than 45 minutes.

4.2 Results

Responses to interpretation questions (16) were coded so that NSpO responses counted as successes (DV = 1) and SpO responses were counted as failures (DV = 0). The data were modeled as various logistic linear mixed effects regression models (Baayen et al., 2008), with dummy coded predictor variables[3] with by-subjects and by-items random slopes and intercepts (Barr et al., 2013). All analyses were conducted within R using the `nlme4` package (Bates and Maechler, 2009) for model fitting. The experimental design permitted numerous measures, such as Responses aggregated across Speakers and Hearers, Speaker response only, Hearer response only, and Percent agreement between Speaker-Hearer pairs, each of which is presented in turn below. Reaction time was not formally examined.

Treating Speaker and Hearer responses as independent events within the same data set – i.e., not distinguishing between Speaker and Hearer responses, Src-Past (M = 82%, SE = 4) elicited significantly more NSpO responses than No-Src (M = 42%, SE = 5), $z = 4.90$, $p < 0.001$, and, in turn, Src-Pres (M = 95%, SE = 2) elicited more NSpO responses than its Src-Past counterpart, $z = 7.33$, $p < 0.001$.[4] The means for each condition

[3] Dummy coding compares each level to a baseline, in this case the No-Src condition; however, qualitatively similar results obtained under ANOVA-style deviation coding, which compares the means of each level against the grand mean.

[4] This is one instance where choice of contrast coding mattered. In ANOVA-style deviation coding, where the No-Src condition was again treated as the baseline for deviation, Src-Pres elicited more non-speaker responses than the grand mean (M = 73%, SE = 2), $z = -8.15$, $p < 0.001$, but Src-Past did not, $t < 1$. However, we concentrate on dummy coding here, as it coheres best with evaluating the predictions against

are shown in Figure 1. Note that the response pattern supports our basic predictions. First, TFRs do not, by themselves, mandate a shift to a Non-speaker commitment. Second, the more cues that are available, the more likely the shift.

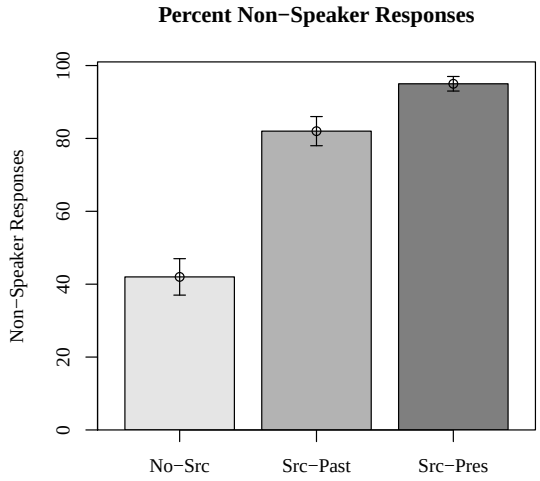


Figure 1: Percent Non-Speaker Responses.

We may also fit the data to a model containing the Role (Speaker, Hearer) of the participant as a predictor (random effect structures were simplified to by-subjects and by-items random intercepts in order allow the model to converge). As before, we find more NSpO responses for Src-Past than the No-Src baseline, $z = 6.27$, $p < 0.001$, and additional NSpO responses for Src-Pres over Src-Past, $z = 6.16$, $p < 0.001$. We also find a small (and possibly spurious) main effect of Role, such that those in the Speaker role (M = 74%, SE = 3) selected NSpO responses more often than those in the Hearer role (M = 71%, SE = 3). This effect is likely to be driven by the 20% increase in the No-Src condition, as there were actually *fewer* NSpO responses for Speakers in the Src-Past condition (d = -10%, with a significant interaction, $z = -2.78$, $p < 0.01$), and no difference whatsoever in the Src-Pres condition, illustrated in Figure 2. At the moment, we do not have a clear account for why participating in different roles may have yielded different behavior in the different sentence types. One possible explanation is that the Speakers failed to produce No-Src sentences with consistent prosody.

One of the benefits of this paradigm is that it provides a measure of Speaker-Hearer agreement. In general, there was a relatively high rate of

the data.

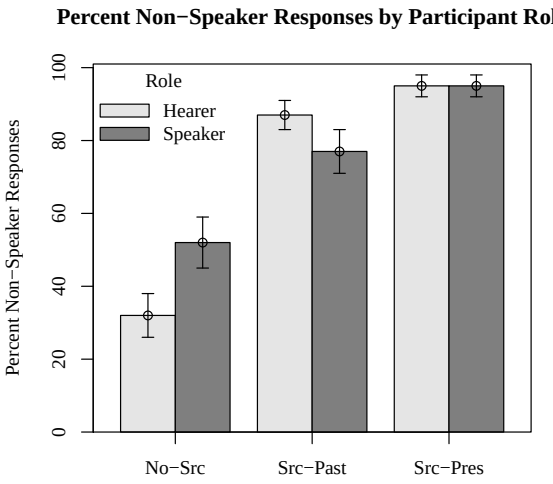


Figure 2: Percent Non-Speaker Responses by Participant Role.

agreement across the entire experiment (including unrelated manipulations) at rate of 62%, significantly above chance in a binomial test, $p < 0.001$. At 73%, the rate of agreement was in fact higher for the present manipulation. Interestingly, participants tended to agree more often on some conditions than others: Src-Pres elicited more agreement (M = 90%, SE = 4) than the No-Src (M = 60%, SE = 6) condition, $z = 3.47$, $p < 0.001$, which did not significantly differ from the Src-Past (M = 70%, SE = 6) condition, $z = 1.25$; see Figure 3.

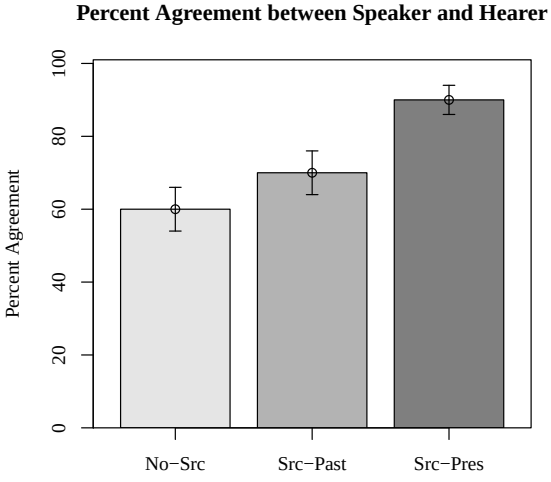


Figure 3: Percent Agreement between Speaker and Hearer.

Further, when participants agreed on the inter-

pretation, the NSpO response was selected at an even greater rate for Src conditions (Pres: M = 100%, SE = 0; Past: M = 95%, SE = 3; this 5% difference between Src conditions was not significant) compared to the No-Src condition (M = 36%, SE = 8), $z = 3.79$, $p < 0.001$.

Although auditory recordings were made of the Speaker's performances, they have not been analyzed in detail. Our impression is that most Speakers simply read the text without giving it much expressive nuance. However, for the few participants who did expressively perform the text, we noted an interesting pattern: Speakers sometimes placed contrastive pitch accent on the source pronoun or the attitude verb, along with a slight pause before the term within the TFR, possibly indicating a quotational effect. We suspect that these intonational cues, among others, would positively correlate with a Non-Speaker interpretation of the TFR. We are currently investigating this issue within a corpus of more natural speech, such as conversations and news reports.

4.3 General discussion

We presented a two-person judgment experiment testing how the presence of a third person source and tense contribute to Non-Speaker interpretations of Transparent Free Relatives. Our findings support the conclusion that TFRs do not semantically signal Non-Speaker interpretations by themselves, as they are consistent with both Speaker and Non-Speaker interpretations. Rather, elements within the TFR serve as subtle, yet reliable, cues for commitment. Specifically, the presence of a Non-Speaker source is a reliable indicator of Non-Speaker commitment, an effect which is increased by the present tense, indicating a habitual, rather than episodic, stance with respect to the attribution described in the TFR. Further, these cues may be used very effectively to signal a shift away from Speaker commitment, as indicated by the high rate of agreement between Speaker and Hearer participants in the experiment.

5 Conclusion

Judgments regarding commitment may not be an all or nothing affair. Hearers rely on subtle pragmatic cues to infer Non-Speaker orientation. Although such interpretations are most likely *invited* inferences, in that they are not mandated by lexical or structural elements, they nevertheless present a crucial component to full comprehension of text and dialogue. This study probed a few factors that give rise to Non-Speaker commitment within the understudied, yet ubiquitous, TFR construction, and showed that various cues work together to strengthen Non-Speaker commitment.

That multiple cues conspire to more effectively indicate Non-Speaker commitment makes intuitive sense. We suspect that deviating from the canonical assumption of Speaker commitment might be a risky endeavor, as the indicators of Non-Speaker commitment are not lexically encoded in English. Should the Speaker fail to successfully communicate her intentions, she runs the risk of being associated with the very point of view from which she wishes to distinguish herself. Thus, using multiple, possibly redundant, cues to cement Non-Speaker interpretations may ensure a greater likelihood of success.

The pragmatics of the TFR construction intuitively parallel issues often discussed in audience design, in that the terms that one uses for an object may reflect a particular conceptualization of that object (Brennan and Clark, 1996). Discourse participants understand that such conceptualizations may well vary, and a conceptual pact to use one mode of reference can be established through continued interaction, in a process called *lexical entrainment*. While the use of TFRs is, in a sense, more general than entrainment in that it applies to more aspects of linguistic communication than copresent reference, we fully expect that common principles govern them both. The case of TFRs is particularly interesting with respect to commitment, as the construction offers a systematic method for pairing a commitment with a source, which is especially important when the term is rich in perspectival information. Nevertheless, TFRs are just one of the many ways that speakers navigate potential disagreement between audience members. We expect that a multitude of cues which discourse participants use to adapt to differing perspectives overlap in the two cases. Understanding how these cues work together will hopefully help us develop more complete models of discourse, along with a richer notion of commitment.

Appendix

Experimental items are provided below. Only the Source-Present condition is given past item 1.

1. John gave Mary what (is called / he called / he calls) a beergarita.
2. Karen made what she calls a goulash.
3. Dylan picked up what he thinks is a rare diamond.
4. Megan ran over what she believes was a mutant rodent.
5. Paterson admitted to what he considers a heinous betrayal.
6. Ken told his boss about what he acknowledges was a grave mistake.
7. The artist sold what she considers her greatest achievement.
8. The television executive promotes what he calls edutainment.
9. The priest performed what he calls a shotgun marriage.
10. The judge condemned the defendant for what he calls a reckless act.
11. The producer released what he expects to be a one hit wonder.
12. The editor denounced what he thinks is a gross abuse of power.

Acknowledgments

This work has benefited greatly from discussions with Lyn Frazier, Chris Potts, and Carson Schütze. Part of this project was presented previously at the UCSD SemBabble group. Many thanks to Ivano Caponigro, Jonathan Cohen, Andrew Kehler and others there for their generous feedback. Thanks also to Mara Breen for her guidance on programming the experiment and to Adrian Staub for use of his lab space and support to run the experiment. Finally, the final version of the paper was improved by following the recommendations of three anonymous reviewers.

References

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Christine Bartels. 1997. *Towards a compositional interpretation of English statement and question intonation*. Ph.D. thesis, University of Massachusetts Amherst.

Douglas Bates and Martin Maechler. 2009. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-31.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Joan Bresnan and Jane Grimshaw. 1978. The syntax of free relatives in English. *Linguistic Inquiry*, 9(3):331–391.

Ivano Caponigro. 2003. *Free not to ask: On the semantics of free relatives and wh-words cross-linguistically*. Ph.D. thesis, University of California, Los Angeles, Los Angeles.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Y Vardi. 1995. *Reasoning about knowledge*. MIT Press, Cambridge, MA.

Donka F. Farkas and Kim B. Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics*, 27(1):81–118, 2.

H. Paul Grice. 1978. Further notes on logic and conversation. Harvard University Press, Cambridge, MA.

Alexander Grosu. 2003. A unified theory of standard and transparent free relatives. *Natural Language & Linguistic Theory*, 21(2):247–331.

Christine Gunlogson. 2001. *True to Form: Rising and Falling Declaratives as Questions in English*. Ph.D. thesis, University of California, Santa Cruz, Santa Cruz.

Christine Gunlogson. 2008. A question of commitment. *Belgian Journal of Linguistics*, 22:101–136.

Charles L. Hamblin. 1970. *Fallacies*. Methuen and Co, London, UK.

Charles L. Hamblin. 1971. Mathematical models of dialogue. *Theoria*, 37(2):130–155.

Jesse A. Harris and Christopher Potts. 2009. Perspective-shifting with appositives and expressives. *Linguistics and Philosophy*, 36(2):523–552.

Jesse A. Harris and Christopher Potts. 2011. Predicting perspectival orientation for appositives. In *Proceedings from the 45th Annual Meeting of the Chicago Linguistic Society*, volume 45, pages 207–221, Chicago, IL. Chicago Linguistic Society.

Jesse A. Harris. 2012. *Processing perspectives*. Ph.D. thesis, University of Massachusetts Amherst, Amherst, MA.

Peter Lasersohn. 2005. Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, 28(6):643–686.

Isaac Levi. 1991. *The fixation of belief and its undoing: Changing beliefs through inquiry*. Cambridge University Press.

Stephen C. Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. The MIT Press, Cambridge, MA.

David Lewis. 1969. *Convention. A philosophical study.* Harvard University Press, Cambridge, MA.

Sophia Malamud and Tamina Stephenson. 2011. Three ways to avoid commitments: Declarative force modifiers in the conversational scoreboard. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 74–83.

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics. Oxford University Press.

Doug Rohde. 2003. Linger. Computer Program.

Carla Schelfhout, Peter-Arno Coppen, and Nelleke Oostdijk. 2004. Transparent free relatives. In *Proceedings of ConSOLE XII*.

Mandy Simons. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056.

Carlota S. Smith. 2003. *Modes of Discourse: The local structure of texts*. Cambridge University Press, Cambridge, UK.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5):701–721.

Chris Wilder. 1998. Transparent free relatives. *ZAS Papers in Linguistics*, 10:191–199.

Helping, I mean assessing psychiatric communication: An application of incremental self-repair detection

Christine Howes[1], **Julian Hough**[1,2], **Matthew Purver**[1] and **Rose McCabe**[3]
[1]Cognitive Science Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London
[2]Dialogue Systems Group, Faculty of Linguistics and Literature, Bielefeld University
[3]Medical School, University of Exeter
c.howes@qmul.ac.uk

Abstract

Self-repair is pervasive in dialogue, and models thereof have long been a focus of research, particularly for disfluency detection in speech recognition and spoken dialogue systems. However, the generality of such models across domains has received little attention. In this paper we investigate the application of an automatic incremental self-repair detection system, STIR, developed on the Switchboard corpus of telephone speech, to a new domain – psychiatric consultations. We find that word-level accuracy is reduced markedly by the differences in annotation schemes and transcription conventions between corpora, which has implications for the generalisability of all repair detection systems. However, overall rates of repair are detected accurately, promising a useful resource for clinical dialogue studies.

1 Introduction

Self-repairs are known to be pervasive in human dialogue and there has been much research into the identification and modelling of repair from both computational and psychological perspectives. In computational linguistics, the focus is on removal of disfluency: for the creation of accurate and useful dialogue systems, disfluencies (including self-repair) need to be identified and removed from the speech input to yield interpretable input for downstream processors (especially when using off-the-shelf parsers). Psycholinguistic research, on the other hand, investigates what the presence and type of repair can tell us about psychological and interactional factors in dialogue. For example, the presence of repair can aid comprehension (Brennan and Schober, 2001) and affect the backchannelling of listeners (Healey et al., 2013). In the psychiatric domain, levels of repair have been found to be associated with verbal hallucinations, and patient adherence to treatment (Leudar et al., 1992; McCabe et al., 2013). Identifying repair in these types of dialogue therefore has the potential to be a diagnostic tool, and offer insights into developing training for psychiatrists, e.g. in detecting that a patient is in difficulty, or shaping their own talk more effectively.

1.1 Self-repair

In the conversation analysis literature (e.g. Schegloff et al. (1977)), repairs are described in terms of the dialogue participant (DP) who initiates the (need for) repair (oneself or another), the DP who completes the repair (self or other), and in which position the repair is completed. For the purposes of this paper, we are interested in cases where a DP repairs their own utterance in the course of producing it – a *position one self-initiated self-repair*, which can repeat part of the utterance (an *articulation* repair, as in (1)), reformulate part of the utterance (a *formulation*, as in (2)), or add something clarificatory to the utterance at a point at which it might have been considered complete (a *transition space* repair (3)).[1]

(1) **Dr:** You probably have seen so many psychiatrists *o- o-* **over** the years

(2) **Dr:** *Did you feel that* **did you despair so much that** you wondered if you could carry on?

(3) **P:** Where I go to do some *printing*. **Lino printing**

Rates of self-repair are known to differ over a startling variety of factors; for example, in different domains and dialogue roles (Colman and

[1]These examples are taken from the psychiatric consultation corpus detailed in Section 2.1, with the reparandum shown in italics and the repair phase shown in bold.

Healey, 2011), modalities (Oviatt, 1995), dialogue moves (Lickley, 2001) gender and age groups (Bortfeld et al., 2001) and clinical populations (Lake et al., 2011). For this reason, there is much discussion in the literature over the underlying cause of self-repair – is it merely an index of difficulty for the speaker, for example when planning or producing an utterance (Bard et al., 2001), or is repair interactively designed for the benefit of the listener(s) (Clark and Fox Tree, 2002; Goodwin, 1979)? While we do not address these questions here, we note that this uncertainty causes repair annotation protocol differences, and makes it unclear whether automatic repair detection trained on any single corpus will generalise to any other.

1.2 Repair in psychiatry

In the psychiatric domain, aspects of doctor-patient communication have been shown to be associated with patient outcomes, in particular patient satisfaction, treatment adherence and health status (Ong et al., 1995). Studies specifically investigating repair show associations between repair and clinical populations known to have language difficulties. For example, Lake et al. (2011) found that participants on the autistic spectrum revised their speech less often than controls, and used fewer filled pauses. For patients with schizophrenia, different rates of repair have been linked to specific types of symptoms, such as verbal hallucinations (Leudar et al., 1992), and whether or not a patient is likely to adhere to their treatment (McCabe et al., 2013) as well as psychiatrist assessments of the therapeutic relationship (McCabe, 2008). These studies rely on accurately hand-annotated repair data, and are not directly comparable to each other as different annotation schemes have been used. Assessing the veracity of these results, and exploring the relationship between repair and outcome – for example, how increased levels of repair are associated with a better therapeutic relationship – requires large datasets to be annotated according to the same schema. This is impractical where expensive and time-consuming hand annotations are required. A domain-general automatic repair identification system would enable us to address some of the specific questions raised by these preliminary results.

1.3 Identifying repair

By hand Self-repairs, which are the repair type of interest in this paper, are often annotated according to a well established structure from (Shriberg, 1994) onwards, and as described in Meteer et al.'s (1995) Switchboard corpus annotation handbook:

$$\underbrace{\text{John and Bill}}_{\text{original utterance}}\ \underbrace{[\ \text{like}}_{\text{reparandum}}\ +\ \underbrace{\{\text{uh}\}}_{\text{interregnum}}\ \underbrace{\text{love }]}_{\text{repair}}\ \underbrace{\text{Mary}}_{\text{continuation}} \tag{4}$$

This structure affords three principal subtypes of self-repairs: *repetitions*, *substitutions* and *deletions*. Repetitions have identical reparandum and repair phases; substitutions have a repair phase that differs from its repair phase lexically but is clearly substitutive of it; and deletions have no obvious repair phase that is substitutive of their reparandum, with utterance-initial deletions often termed *restarts*. Despite the clarity the structure affords, there is often low agreement between annotators deciding between substitutions and deletions; in fact, considering gradient boundaries between these categories may be more useful (Hough and Purver, 2013). Presence of a repair alone is agreed upon more often than structure.

While this annotation scheme has been widely used in the computational linguistics community, this is not as common for repair corpus studies interested in the dialogue function of repair, rather than their surface structure. Healey et al. (2005) present a systematic effort to test the reliability of a human annotation scheme for repair, building on Healey and Thirlwell's (2002) annotation protocol for identifying the different CA types of repair in dialogue transcripts. They divide repairs into the CA categories of Position 1 repair (*Articulation*, *Formulation*, *Transition space* as shown in (1)-(3), above), Position 2 repair (*Clarification Request/NTRI*, *Correction*) and Position 3 (*Follow-up* and *reformulate*). Healey et al. (2005) tested the validity and reliability of the protocol through an analysis of two of the authors coding a corpus of repair sequences drawn from the CA repair literature with their original coding removed. The validity of the protocol was shown to be encouraging overall, with 75% of the repairs being assigned the same category as that of the original papers, though detection agreement rates were not reported.

Automatically There has been considerable work on detecting reparandum words from transcripts, with the motivation of filtering them out before parsing. However, while the computational linguistics community focusses on the Switchboard corpus disfluency challenge (Charniak and Johnson, 2001), which has been met with considerable success in terms of reparandum word detection (Honnibal and Johnson, 2014; Rasooli and Tetreault, 2014), these models have rarely been applied outside of this domain. This is because there is a lack of gold-standard disfluency annotation in the format shown in (4) available: in fact, Switchboard provides the only large consistently annotated corpus available for this purpose. Furthermore, the fine-grained utterance unit segmentation as carried out by the Switchboard disfluency scheme (Meteer et al., 1995) is uncommon in other corpus mark-ups. For this reason, cross-domain efforts have been rare and performance dips considerably across domains (Lease et al., 2006; Zwarts et al., 2010b). Furthermore, such models are often not designed with word-by-word incremental processing (as required in an incremental dialogue system) in mind; the only effort to develop a system that could function incrementally in a reliable way (Zwarts et al., 2010a) suffers from latency issues, not detecting repairs until an average of 4.6 words after the repair onset.

While the fine-grained structural detection of repairs is necessarily the focus in computational work, to allow reconstruction of a "cleaned" utterance, high accuracy on detecting the structure may be unnecessary for tasks focussing on inter-subjective *rates* of repair. Use of gold-standard Switchboard-style repair annotations in supervised machine learning approaches has a tendency to cause tight fitting to the Switchboard annotation and transcription conventions. While this data can be used as a basis to train a system, it needs to be suitably adaptable to different corpora.

1.4 Research questions

This study applies an incremental repair detection system (STIR; see Section 2.2, below) trained and initially tested on the Switchboard corpus, to a corpus of face-to-face clinical dialogues between patients with schizophrenia and their psychiatrists. The questions we are directly concerned with are:

- Can self-repair be consistently detected across domains and modalities?
- How reliably can different annotation schemes for repair be compared?
- How useful is automatic analysis of self-repair in the clinical domain?

2 Methods

2.1 Data

Switchboard The Switchboard disfluency tagged corpus (Godfrey et al., 1992; Meteer et al., 1995) which has Penn Tree Bank III mark-up, consists of 650 dyadic telephone conversations collected between 1990 and 1992 between unfamiliar American participants on a range of topics assigned from a pre-determined list, ranging from 1.5 up to 10 minutes in duration, with the average conversation lasting around 6.5 minutes. The disfluencies annotated include filled pauses, discourse markers, and edit terms, all with standardised spelling e.g. consistent 'uh' and 'uh-huh' orthography. First-position self-repairs are bracketed with the structure in (4) with reparandum, interregnum and repair phases marked. It has gold standard Penn Tree Bank part-of-speech (POS) tags and is segmented in terms of sub-turn utterance units. Restart repairs (utterance-initial deletions) are coded as two separate units and not in fact annotated as repairs.

Psychiatric consultation corpus (PCC) The clinical corpus was constructed using a subset of data from a study investigating clinical encounters in psychosis (McCabe et al., 2013), collected between March 2006 and January 2008. The corpus consists of transcripts from 51 outpatient consultations of patients with schizophrenia and their psychiatrist. These transcripts relate to 51 different patients, and 17 psychiatrists. The consultations varied in length, with the shortest consisting of only 709 words (lasting approximately 5 minutes), and the longest 8526 (lasting nearly an hour). The mean length of consultation was 3500 words.

Each transcript was hand-annotated for repair using the protocol described in Healey et al. (2005). For each turn, words in repairs and their reparanda were highlighted using Dexter Coder (Garretson, 2006). The resulting annotations are available in a standalone XML format. For the purposes of this study, the data extracted consisted of the transcripts and associated position 1 repairs (annotated with reparandum phrase and corresponding repair phase). Filled pauses are not

explicitly annotated, but are identifiable as interregna as the unannotated text between the end of the reparanda and its repair. Filled pauses, while consistently transcribed, were found to be inconsistently spelt (*aammm*, *er*, *eerrrrmm*, *uhmmm* etc). A find-and-replace operation was therefore applied to the corpus prior to analysis to give these a standardised spelling, i.e. a consistent 'er'. Prior to the analysis, we also tagged the corpus for part-of-speech using the Stanford POS tagger (Toutanova et al., 2003). The Stanford tagger is trained on written text, and previous work applying it to spoken dialogue has shown the error rates to be in the order of 10% (Mieskes and Strube, 2006). Here, we are not concerned with the POS labels per se, but in the parallelism between POS label sequences (see below) - given that errors are likely to be fairly consistent (dependent on transcription spelling or spoken dialogue idiosyncracies) we take this as sufficient for our purposes.

2.2 STIR: Strongly incremental repair detection

As a repair detection framework we use the STIR (STrongly Incremental Repair detection) system, designed with incrementality and domain-generality in mind (see Hough and Purver (2014)). STIR does not require much annotated disfluency data to become practically useful, as its backbone is derived from simple language model features. Additionally, due to its pipelined classifier structure, different phases of the repair structure in (5) can be included or excluded, depending on the detection task and the available annotations. The repair structure in (5) maps directly to that shown in (4), with the start and end word of the reparandum marked by rm_{start} and rm_{end}, the optional interregnum marked as ed and the repair phase delimited by rp_{start} and rp_{end}.

$$...[rm_{start}...rm_{end} + \{ed\}rp_{start}...rp_{end}]... \quad (5)$$

STIR's pipeline structure is intended to support incremental processing while being cognitively plausible: it first detects edit terms ed (where present), and then the repair onset rp_{start}; subsequent stages then identify the extent of the reparandum rm_{start} and the end of the repair rp_{end}. Here, we are interested only in repair points, so use only the first two steps – for full details see Hough and Purver (2014).

2.2.1 Enriched language models

STIR is driven by probabilistic models of language which approximate *fluency* level. This is in contrast to most machine learning approaches to repair tagging which often use string alignment for repeated words and POS tags as their principal features. This allows STIR to be compatible with annotation protocols such as (Healey et al., 2005; Colman and Healey, 2011) more concerned with the rate, dialogue type and presence of disfluency rather than purely for identifying reparanda. STIR can thus be used for different repair detection tasks, adapting to the available annotations, and the motivations for the repair detection.

Following Hough and Purver (2013), STIR uses enriched Kneser-Ney (Kneser and Ney, 1995) smoothed trigram language models, trained on a corpus *with disfluencies removed*. The most basic fluency feature is the negative log of the smoothed trigram probability value s (equation 6), aka the *surprisal*. We also use features that approximate syntactic fluency, the principal measure being the (unigram) Weighted Mean Log probability (WML) of utterances and their local trigrams (equation 7), a feature that factors out the contribution of lexical rarity. WML was originally used successfully in detecting low grammaticality judgements (Clark et al., 2013) and given the word-by-word Markov independence assumption of n-gram models it serves as an approximation of incremental syntactic fluency.

$$\mathrm{s}(\mathrm{w}_{i-2} \dots w_i) = -\log_2 p^{kn}(w_i \mid w_{i-2}, w_{i-1}) \quad (6)$$

$$\mathrm{WML}(\mathrm{w}_{i-2} \dots w_i) = \frac{\log_2 p^{kn}_{TRIGRAM}(\langle w_{i-2}...w_i \rangle)}{-\sum \log_2 p^{kn}_{UNIGRAM}(\langle w_{i-2}...w_i \rangle)} \quad (7)$$

These feature values can now be calculated at each word, with versions based on word (s^{lex}, WML^{lex}) and POS tag (s^{POS}, WML^{POS}) sequence. For the WML values, we also calculate the difference between values at current and previous word/POS (ΔWML). This gives 6 features overall.

2.2.2 Additional features

STIR's classifiers combine these language model features with further specific logical (binary) features. The *alignment* features indicate whether the word/POS W_x in position x in a trigram is identical to the final word/POS in the trigram, $W3$. The *edit* feature is true iff there is an edit term (filled pause, edit term or discourse marker) detected in the position before $W3$ – see Table 1.

Word n-gram features (n=3)	s^{lex}, WML^{lex}, ΔWML^{lex}
POS n-gram features (n=3)	s^{POS}, WML^{POS}, ΔWML^{POS}
Alignment features (n=4)	$W2 = W3$, $W1 = W3$, $POS2 = POS3$, $POS1 = POS3$
Edit term feature (n=1)	$edit$ [1,0]

Table 1: Repair onset detection features

2.2.3 Training and testing

For the 6 language model features, we train word and POS 'fluent' language models on the standard Switchboard training data (all files with conversation numbers beginning sw2*, sw3* in the Penn Treebank III release), consisting of ≈100K utterances, ≈600K words, cleaned of disfluencies (i.e. edit terms and reparanda) and with gold-standard POS tags. We then keep this language model the same when calculating the feature values across different test corpora; these consist of raw dialogue transcripts with disfluencies included. When testing on data other than Switchboard, the POS tags are generated using the Stanford POS tagger (see above).

As the test corpora have disfluencies present, partial words may be present, either explicitly transcribed as such, or detected by observing an unknown word that forms an orthographic prefix of its following word (i.e. 's, so'). As corpus studies suggest that a non-utterance-final partial word presence predicts a disfluency almost perfectly, for multi-word as well as single partial-word disfluent cut-offs (Hough and Purver, 2013), we include them into STIR's language models with a probabilistic penalty (see Hough and Purver (2014) for details).

Edit term detection uses the word and POS n-gram features above, plus the likelihood assigned by an edit term language model derived from Switchboard's training data. After edit word detection, for repair detection we obtain values for the features listed above for each of the remaining words in a word-by-word fashion from the standardly used Switchboard heldout data (files PTB III sw4[5-9]*; 6.4K utterances, 49K words).

2.2.4 Classifier pipeline

STIR's first two stages are then implemented as random forest classifiers (Breiman, 2001): the first classifies whether the last word seen is an edit term (ed) or not, and the second classifies whether the word is a repair onset (rp_{start}) or not. If the ed classifier classifies a word as ed, the word is not considered for rp_{start} classification; consequently edit term detection is the first stage in the disfluency detection pipeline. We employ weighted error functions to balance recall and precision in the desired way for the detection task using MetaCost (Domingos, 1999). This allows fine-grained control over the rate of onset prediction, which proved to be very useful for the clinical data.

2.3 Experimental set-up

We choose the cost functions for MetaCost on Switchboard heldout data to yield the best overall F-score of rp_{start} detection, we then test on the test data on the standard Switchboard test files (PTB III sw4154 - sw4483; 6.7K utterances, 48K words) for the precision, recall and F-scores and 'relaxed' repair-per-turn evaluation of repair detection (see below for details). For the PCC data, while we keep the base classifier the same as Switchboard, we optimise the weights to balance precision and recall on a heldout set of doctor-patient interaction of ≈20K words. This step was carried out as the weights used for Switchboard yielded much higher precision than recall in rp_{start} detection on a word-by-word level, though the overall accuracy was roughly the same. We then test on a different set of ≈25K words.

3 Results and discussion

Edit term detection Edit term detection was evaluated on the Switchboard test data, achieving an F-score of 0.938. While this is not directly comparable to previous work, Heeman and Allen (1999) also report very high accuracy on detecting a subset of edit terms, *discourse markers*, achieving an F-score of ≈0.96. Our system detects a bigger and more variable class of phenomena.

Testing edit-term detection on the PCC data was more difficult, as edit terms were not explicitly annotated. For the PCC data, transcribed filled pauses are automatically tagged as edit terms and then edit term detection is performed using a model trained on the Switchboard data – this serves as an approximation only; due to the lack

of gold standard this was not evaluated quantitatively, but see below for discussion.

Repair point detection We then tested STIR in terms of its precision, recall and F-score for repair onset detection as in (8).

$$\begin{aligned} \text{precision} &= \frac{rp_{start}\ \text{correct}}{rp_{start}\ \text{hypothesised}} \\ \text{recall} &= \frac{rp_{start}\ \text{correct}}{rp_{start}\ \text{gold}} \\ \text{F-score} &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \end{aligned} \tag{8}$$

We evaluate in two ways: a *strict* evaluation at the word level, requiring the exact repair point word rp_{start} to be identified; and a *relaxed* evaluation at the turn level, with a rp_{start} hypothesis taken as correct if in the same turn as a gold-standard repair annotation, but with every additional hypothesised rp_{start} over the correct number treated as a false positive (i.e. incrementing rp_{start} hypothesised but not rp_{start} correct). The results are shown in Tables 2 and 3.

detection	precision	recall	F-score
strict	0.862	0.755	0.805
relaxed	0.904	0.787	0.841

Table 2: Switchboard test data results

detection	precision	recall	F-score
strict	0.527	0.536	0.532
relaxed	0.682	0.679	0.680

Table 3: PCC test data results

Turn-level data As can be seen in Table 2, on the Switchboard data the system identifies both that there is a repair and its exact position in the turn very well (F-score > 0.8). However, for the PCC data (see Table 3), although the system identifies that there are repairs in the turn reasonably well (F-score ≈ 0.7), there is a large drop in performance when looking at the strict position-based metric (F-score ≈ 0.5).

This is likely to be due to differences in both transcription and annotation conventions. In the PCC data, the emphasis for annotators was on identifying the number and type of repairs in the turn. Although there was good agreement between annotators at this level – with levels comparable to our relaxed evaluation performance (Cohen's κ = 0.73, (McCabe et al., 2013)), it is not clear whether the annotators position repair points systematically or agree on positioning. Examination of the transcripts suggests that annotation differences can abound. For example, as shown in (9), editing phrases such as 'I mean' may be annotated as part of the reparandum (9a), left unannotated between reparandum and repair (9b), or annotated as part of the repair itself (9c). While (9b) maps most directly to the Switchboard annotation schema, these differences do not affect the overall number and type of repairs found in a turn, and are therefore only relevant if our task is the strict one of finding the exact position of repairs. While this is usually important for the purposes of speech recognition or dialogue systems, it is not here – our interest in is the association between outcomes and the presence and rate of different types of repairs.

(9) **(a) Dr:** well *I think I mean* **I think** that's why it's really sensible

(b) Dr: well *I think* I mean **I think** that's why it's really sensible

(c) Dr: well *I think* **I mean I think** that's why it's really sensible

Dialogue level data Given the differences in turn-level data, as outlined above, and the different ways in which automatically annotated repair data might be used, we compared the number of identified repairs over each dialogue.

As can be seen from Table 4, there is a very high correlation (> 0.9) between the number of repairs per transcript detected by the automatic incremental classifier and those annotated by hand. At this coarse-grained level, the system provides a useful overview of self-repair, which can allow us to make comparisons between speakers who typically use a lot of repair and those who do not, as well as looking for associations with outcomes on a by-patient level as in (McCabe et al., 2013). However, as can also be seen in Table 4, the automatic repair numbers are lower than those for the hand-coded data, and this is especially the case where patients are concerned. This indicates that the system is systematically *not* picking up certain types of repair that the patients are using.

When comparing the hand annotations on the PCC data with STIR's output, we see differences

	Hand-coded		Automatic		Correlation	
	Mean	(s.d.)	Mean	(s.d.)	r	p
Patient P1 repair	62.51	(44.87)	48.90	(33.29)	0.945	< 0.001
Doctor P1 repair	41.57	(23.25)	41.02	(23.23)	0.906	< 0.001

Table 4: Relationship between hand-coded and automatically generated repair measures

due to several factors of annotation protocol and behaviour and not just due to inherently poor system performance. See examples (10)-(12) where the hand annotation tags (shown in (a) in each case) differ from STIR's annotations (shown in (b)).

(10) **(a) D:** ... and if you tell me that **that**$[RP_{START}]$ that the depressions kicks in ...

(b) D: ... and if you tell me that **that**$[rp_{start}]$ **that**$[rp_{start}]$ the depressions kicks in ...

(11) **(a) D:** and so **I**$[RP_{START}]$ mean otherwise I'm not too concerned about your mental health...

(b) D: and so **I**$[ed]$ **mean**$[ed]$ otherwise I'm not too concerned about your mental health...

(12) **(a) P:** I don't **I'm**$[RP_{START}]$ not like hearing voices...

(b) P: I don't I'm not like hearing voices...

In (10) the second repeat of 'that' is evaluated as a false positive by STIR, reflecting the embedded repairs often found in Switchboard, while the annotator views this as part of one longer repair. A false negative from STIR can be seen in (11) where an annotator deems this a repair, while according to Switchboard, and STIR, this would be an editing phrase 'I mean'. In (12), another false negative is evaluated as STIR misses the transcribed repair onset from 'I'm not'. Utterance-initial deletions, or 'restarts', are not marked in Switchboard but treated as two separate utterance units, so there is no training data for these types of self-repair.

4 Towards domain-general repair detection

Using a more strict word-by-word evaluation, we saw that the differences in annotation schemes and transcription conventions have a marked effect on the system's performance. Switchboard annotation conventions result in a biasing on particular types of repair, namely, mid-utterance repetitions, deletions and substitutions, whereas it is not marked for restarts, which caused it to perform poorly on detecting them in the clinical data. On the clinical side, the fact that editing terms are often marked as the repair onset means a Switchboard-trained detector will not get the exact position of the repair. This has implications for the generalisability of all repair detection systems that rely on strict word-by-word evaluation, such as those used in dialogue systems – the way in which the training data has been annotated and transcribed will affect what types of repair it reliably detects.

Despite the differences in the type of disfluency annotation available, one can build a system that is practically useful for detection purposes using the set-up as shown in Figure 1. As long as there is some heldout data available of the same type as the target corpus, even if not considerable in size, STIR's error functions can be manually adjusted (or automatically experimented with) to yield the best accuracy results before testing. This technique is effective in terms of giving results with good overall correlations as described above.

The element of Figure 1 not present in the version of STIR here is the *"fluent" corpus* which could form additional training data to the fluent language model in STIR. We hypothesize that the appropriate data, even if from written, rather than spoken sources, could boost results on out-of-domain (non-Switchboard) data. (Zwarts and Johnson, 2011) show how large text-based corpora included in a repair hypotheses re-ranker can improve detection on Switchboard, however we would like to explore the effect of additional resources in improving performance on other data, such as the PCC corpus described here. Other data STIR does not currently use is acoustic information, which has been shown to help disfluency detection (Liu et al., 2003). Incorporating speech signal information will form part of future work.

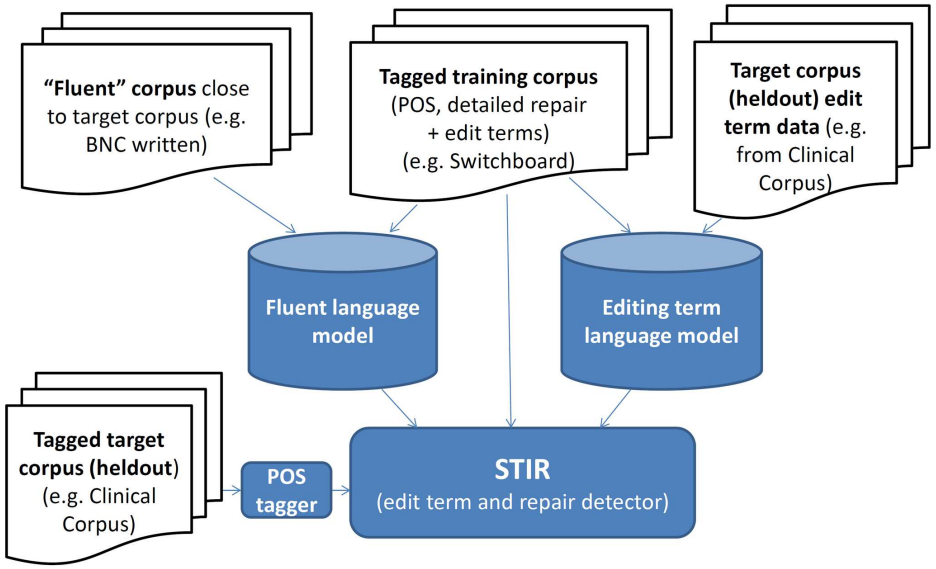


Figure 1: STIR training and heldout sources for a new target domain

5 Conclusions

In terms of the research questions set out in section 1.4, we can detect self-repair reliably across modalities and domains, but only if we use a relaxed evaluation metric. However, this is sufficient for the purposes of examining overall rates of repair, as used in some clinical studies (McCabe et al., 2013), and automatic self-repair detection using STIR can therefore be usefully applied to these datasets, removing the need for time-consuming and costly hand annotations.

The STIR system is intended to provide a domain-general incremental repair detection system and we are currently experimenting with different language models that allow it to generalise to other data in very different dialogue domains. Issues to consider in future work that have been raised by this preliminary study include (but are by no means limited to) the transcription of filled pauses and overlapping speech, how turns are segmented, and issues arising due to the lack of gold-standard POS tags– joint POS-tagger and repair detection could lead to a more robust final outcome (Heeman and Allen, 1999).

In terms of practical applications, the STIR system is already being used to look at changes in self-repair behaviours before and after training in a psychiatrist communications study, and as it is strictly incremental, it has the capacity to be implemented in artificial mental health worker dialogue agents (Faust and Artstein, 2013).

Acknowledgments

Thanks to the three anonymous SemDial reviewers for their helpful comments.

Howes was supported by the EPSRC-funded PPAT project grant number EP/J501360/1 during this work.

Hough is supported by the DUEL project financially supported by the Agence Nationale de la Research (grant number ANR-13-FRAL-0001) and the Deutsche Forschungsgemainschaft. Much of the work was carried out under an EPSRC DTA scholarship at Queen Mary University of London.

Purver is partly supported by ConCreTe: the project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

The clinical data was collected and transcribed as part of the MRC funded project "Doctor-patient communication in the treatment of schizophrenia: Is it related to treatment outcome?" (G0401323).

References

Ellen G. Bard, Robin J. Lickley, and Matthew P. Aylett. 2001. Is disfluency just difficulty? In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.

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.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

S.E. Brennan and M.F. 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 North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–9. Association for Computational Linguistics.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111.

Alexander Clark, Gianluca Giorgolo, and Shalom Lappin. 2013. Statistical representation of grammaticality judgements: the limits of n-gram models. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 28–36, Sofia, Bulgaria, August. Association for Computational Linguistics.

M. Colman and P. G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.

Pedro Domingos. 1999. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164. ACM.

Lauren Faust and Ron Artstein. 2013. People hesitate more, talk less to virtual interviewers than to human interviewers. In *Proceedings of the 17th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialDam)*, pages 35–43, Amsterdam, dec.

Gregory Garretson. 2006. Dexter: Free tools for analyzing texts. In *Actas de V Congreso Internacional AELFE*, pages 659–665.

John J. Godfrey, Edward Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE ICASSP-92*, pages 517–520, San Francisco, CA.

Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. In G. Psathas, editor, *Everyday Language: Studies in Ethnomethodology*, pages 97–121. Irvington Publishers, New York.

P. G. T. Healey and M. Thirlwell. 2002. Analysing multi-modal communication: Repair-based measures of communicative co-ordination. In *Proceedings of the International CLASS Workshop on Natrual, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, pages 83–92, June 28th-.

P. G. T. Healey, M. Colman, and M. Thirlwell. 2005. Analysing multi-modal communication: Repair-based measures of human communicative co-ordination. In J. van Kuppevelt, L. Dybkjaer, and N. Bernsen, editors, *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, volume 30 of *Text, Speech and Language Technology*, pages 113–129. Kluwer, Dordrecht.

Patrick G. T. Healey, Mary Lavelle, Christine Howes, Stuart Battersby, and Rose McCabe. 2013. How listeners respond to speaker's troubles. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, Berlin, July.

Peter Heeman and James Allen. 1999. Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

Matthew Honnibal and Mark Johnson. 2014. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association of Computational Linugistics (TACL)*, 2:131–142.

Julian Hough and Matthew Purver. 2013. Modelling expectation in the self-repair processing of annotat-, um, listeners. In *Proceedings of the 17th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialDam)*, pages 92–101, Amsterdam, December.

Julian Hough and Matthew Purver. 2014. Strongly incremental repair detection. In *Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. (to Appear).

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Johanna K Lake, Karin R Humphreys, and Shannon Cardy. 2011. Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders. *Psychonomic bulletin & review*, 18(1):135–140.

Matthew Lease, Mark Johnson, and Eugene Charniak. 2006. Recognizing disfluencies in conversational speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1566–1573.

Ivan Leudar, Philip Thomas, and Margaret Johnston. 1992. Self-repair in dialogues of schizophrenics: Effects of hallucinations and negative symptoms. *Brain and Language*, 43(3):487 – 511.

Robin J Lickley. 2001. Dialogue moves and disfluency rates. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.

Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. In *Preceedings of Eurospeech*, pages 957–960.

R. McCabe, P. G. T. Healey, S. Priebe, M. Lavelle, D. Dodwell, R. Laugharne, A. Snell, and S. Bremner. 2013. Shared understanding in psychiatrist-patient communication: Association with treatment adherence in schizophrenia. *Patient Education and Counselling*.

R. McCabe. 2008. Doctor-patient communication in the treatment of schizophrenia: Is it related to treatment outcome? Technical report, Final report on G0401323 to Medical Research Council.

M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer. 1995. Disfluency annotation stylebook for the switchboard corpus. ms. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Margot Mieskes and Michael Strube. 2006. Part-of-speech tagging of transcribed speech. In *Proceedings of LREC*, pages 935–938.

L.M.L. Ong, J.C.J.M. De Haes, A.M. Hoos, and F.B. Lammes. 1995. Doctor-patient communication: a review of the literature. *Social science & medicine*, 40(7):903–918.

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

Mohammad Sadegh Rasooli and Joel Tetreault. 2014. Non-monotonic parsing of fluent umm I mean disfluent sentences. *EACL 2014*, pages 48–53.

E.A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259.

Simon Zwarts and Mark Johnson. 2011. The impact of language models and loss functions on repair disfluency detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 703–711, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Zwarts, Mark Johnson, and Robert Dale. 2010a. Detecting speech repairs incrementally using a noisy channel approach. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1371–1378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Zwarts, Mark Johnson, and Robert Dale. 2010b. Repurposing corpora for speech repair detection: Two experiments. In *Australasian Language Technology Association Workshop 2010*, page 99.

Analysis of the Responses to System-Initiated Off-Activity Talk in Human-Robot Interaction with Diabetic Children*

Ivana Kruijff-Korbayová
Stefania Racioppa
Bernd Kiefer
DFKI, Saarbrücken, Germany
ivana.kruijff@dfki.de

Elettra Oleari
Clara Pozzi
Alberto Sanna
FCSR, Milan, Italy
oleari.elettra@hsr.it

Abstract

We carried out an exploratory WOZ study with a conversational human-robot interaction system which offers a set of activities aimed to help a child to improve its capability to manage diabetes. The novel aspect is the inclusion of robot-initiated *off-activity talk* (OAT) on diabetes- and health-related topics. We present an analysis of the OAT sub-dialogues: their distribution, the prompts, children's responses, engagement. Children generally engaged well. They sometimes also reciprocated the robot's topics and even took initiative with new ones. On the other hand, we observed a decline in children's engagement as the interactions progressed. We attribute this mostly to the delays in system response, due to the WOZ setup.

1 Introduction

The work presented here is part of the ALIZ-E project (Aliz-E, 2014). We investigate the use of a robotic companion to provide support to diabetic children, who need to acquire knowledge about diabetes and suitable healthy nutrition, develop various relevant skills and learn to adhere to the therapy requirements, in order to become able to manage their condition themselves (Nalin et al., 2012; Belpaeme et al., 2013).

The system developed in ALIZ-E uses the Nao robot (Aldebaran, 2014) to engage a child in several different activities (cf. §3). Since previous research has established that social aspects of interaction are important to sustain long-term engagement of humans with artificial agents, including both virtual characters and robots (cf. §2), the interactions with the ALIZ-E system include both *activity talk*, i.e., conversation pertaining to the activity at hand, and *social talk*, such as greetings and personal introductions.

*This work was funded by the EU FP7 project Aliz-E, grant No. ICT-248116. URL: www.aliz-e.org

The novel aspect in the present explorative study is the inclusion of *off-activity talk* (OAT). Interspersed within activity talk, but not pertaining directly to the activity at hand, OAT involves discussion of diabetes- and health-related topics with the aim to elicit talk from the child, in particular, to encourage disclosure of personal habits and experiences. If successful, OAT could provide a therapeutically valuable instrument to help the doctors and nutritionists to monitor the children's behaviors and hopefully also to motivate the children to adhere to specific therapy-related requirements.

To investigate the viability and impact of OAT and collect empirical data we carried out an experiment during a summer camp for diabetic children (cf. §3). In (Kruijff-Korbayová et al., 2014) we describe the experiment in detail and preent an analysis of children's perception of and relationship to the robot, interest in further interaction(s) and adherence to therapy-related requirements, namely filling a nutritional diary during the summer camp. In the present paper we focus on OAT: its design (§4) and the experience with it (§5 and §6).

2 Background

(Bickmore and Picard, 2005) coined the term *relational agents* for computational artifacts designed to establish and maintain long-term social-emotional relationships with their users. Their team carried out numerous pioneering studies to evaluate the effects of various aspects of (virtual) agent behavior on long-term engagement, e.g., (Bickmore et al., 2010). Relational behavior strategies are also investigated in human-robot interaction, e.g., robots as companions (Lee et al., 2006; Chidambaram et al., 2012; Adam et al., 2010; Nalin et al., 2012) or in therapeutic and educational settings (Kanda et al., 2004; Dautenhahn et al., 2005; Kidd and Breazeal, 2007; Fasola and Mataric, 2012).

It is often underlined that to build long-term

bonds with (young) users and provide them support and motivation, a robot needs to be able to sustain *social dialogues*, including abilities like initial greetings, chatting, and expressing personal opinions and beliefs (Higashinaka et al., 2010). Initial greeting, in particular, is a social skill which (Kahn et al., 2008) considered one of the eight most important design patterns in human robot interaction. Moreover self-disclosure and empathy can contribute to familiarity between two agents engaged in a conversation (Reis and Shaver, 1998; Moon, 2000).

(Bickmore and Cassell, 2001) were the first to use an explicit dynamically updated model of the agent-user relationship. Their social dialogue planner was designed to sequence agent task and small talk utterances to satisfy both task and relational constraints. Several other virtual agents with hand-crafted small talk dialogue strategies are overviewed in (Klüwer, 2011), who proposes a functionally-motivated taxonomy of small talk dialogue acts based on the social science theory of face and extracted dialogue act sequences for social talk from an annotated corpus. (Adam et al., 2010) on the other hand, analyzed a corpus of child-adult conversations to extract so-called personalization behaviors. They identified strategies for gathering and exploitation of personal information (e.g. family, friends, pets); preferences (e.g. favorite movie, favorite food); agenda (plays football on Saturday, has maths every Thursday); activity-specific information (preferred stories, current level of quiz difficulty); interaction environment (e.g. time, day, season, weather).

Small talk is similar in structure to OAT. However, OAT has the purpose to encourage the child's self-disclosure on topics in the domain of diabetes- and health-related concepts. In the area of healthcare and education there is growing body of research on systems to interview patients and consumers about their health and provide health information and counseling using natural language dialog (Bickmore and Giorgino, 2006). Such dialogues have similar content as OAT. In our system we are using game-like activities as a context within which OAT takes place.

3 Experiment

The data analyzed in this paper was collected during the experimental activities described in detail in (Kruijff-Korbayová et al., 2014), carried out in August 2013 at a Summer Camp for diabetic children organized in Misano Adriatico (Italy) by the Center for Pediatric and Adolescent Endocrinology of San Raffaele Hospital (Milan) in cooperation with the Italian patients association SOStegno70 (Sostegno70, 2014).

3.1 Participants

In total 62 children (age 11-14) attended the summer camp and were exposed to the Nao robot (Aldebaran, 2014) during various joint activities. 24 children volunteered to participate in individual session(s) with the robot. 13 of them (7 male, 6 female) were randomly assigned to the OAT condition of interaction. In this paper we analyze the dialogue data collected with these children.

3.2 System

The interactions were carried out using the system developed in the ALIZ-E project (Belpaeme et al., 2013), in a partial Wizard-of-Oz setup. The following activities were available: (i) Quiz, in which the child and the robot ask each other series of multiple-choice quiz questions from various domains (Kruijff-Korbayová et al., 2012a); (ii) SandTray, where the robot and the child solve sorting tasks on a shared touch-table (Baxter et al., 2012); (iii) Dance, where the robot explores various moves with the child, making a connection between motions and nutritional concepts (Ros et al., 2011; Ros et al., 2014). Fig. 1 shows children performing the activities and the room with the experimental setup.

One and the same wizard operated the system in all interactions, and was supervised by a psychologist. The wizard simulated the recognition and interpretation of the user's speech[1] and for OAT also the next system action selection. We provided an interface for the wizard to trigger OAT: The wizard thus could select an OAT dialogue move as the next system action from a set of given options at any point during an activity. The verbalization was done automatically or the wizard could type something in on the fly. The next system action in the Quiz, Dance and SandTray 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

[1] We did not introduce any noise into the child input to simulate speech recognition errors in this experiment.

Figure 1: Left to right: The experimental setup during the summer camp and children engaged in activities with the ALIZ-E system: dance, quiz, sandtray. (anonymized)

an italian voice developed in the project (Kruijff-Korbayová et al., 2012b). Spoken output verbalization was designed so as to ensure high degree of variation in the system output (Kruijff-Korbayová et al., 2012b).

3.3 Procedure

Each volunteer child had a scheduled appointment in their spare time during the day. Before the session, the child was informed about the experiment, instructed about the system and the available activities and filled in a demographic questionnaire.

After this initial phase the interaction started. The robot introduced itself with its name, and asked the name of the child. It then explained the rules and they started to play, first the Quiz game. The children were then free to switch between the three activities and to stop the game at any time. If not previously interrupetd by the child, the session ended after 30 minutes of continuous interaction.

After the interaction, the child was debriefed and could make an appointment for another session with the robot.

We made video and separate audio recordings.

4 Off-Activity Talk Design

The following OAT topics were defined in strict collaboration with a psychologist of the San Raffaele Hospital:

- 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
- Nutritional diary: function; filling in; motivation

We formulated several OAT prompts for each topic and implemented them as canned text utterances in the system, as illustrated below:

- Hobbies: *What do you like to do in your spare time?* or *Do you do any sport or another activity?*
- Diabetes: *Do you inject insulin yourself?* or *If your glycemia is low, what do you do?*
- Nutrition: *How often do you eat fruit and vegetables?* or *What are you favorite foods?*
- Friends: *Do your friends know about diabetes?* or *When you go out, do you take your glucometer and insulin?*
- Adults: *How do your parents behave with you with respect to diabetes?*
- Nutritional diary: *Can you explain to me how the diary works?* or *Is it difficult to fill in the diary?* or *I guess it's difficult but it is very important and useful to do so.*

In Quiz OAT is triggered between question-answer sequences. The first step for the robot to start OAT is to say something to "escape" from the Quiz talk, e.g., *Now, I am curious about something.* The next step is to raise one of the topics as illustrated above. OAT on a given topic can continue by additional utterances in order to create a more complex extended sub-dialogue. Finally, the Quiz activity is resumed explicitly by saying, e.g., *OK, now let's do another quiz question.*

In Dance we defined several OAT utterances to be interlaced with the sequence of movements and sounds, and triggered when the robot begins to explain the related nutritional concepts. Similarly to Quiz, the Dance activity would be explicitly interupted for OAT and resumed afterwards.

In SandTray OAT about nutritional habits can be triggered while the child is playing a sorting game about food and carbohydrates. The game structure makes it easy to raise OAT topics related to the object shown on the tablet, e.g. asking *What food do you prefer between these?* or *Is there any food among these that you put in your food diary?*. OAT thus usually does not need to interrupt the SandTray activity, and there usually need not be an "escape" turn like in Quiz. Consequently, it is also not necessary to explicitly resume the activity.

ID	#topics	#subdial.	#turns R	#turns C	#init. C
02	3	12	33	35	4
04	3	3	25	26	4
05	4	8	25	23	1
09	4	11	20	20	1
10	2	4	22	22	-
11	5	13	25	22	1
12	5	12	24	21	2
14	3	7	16	7	-
17	2	3	19	18	4
18	3	9	25	21	-
19	1	3	7	4	-
22	2	5	17	15	2
24	4	13	37	25	-

Table 1: Distribution of OAT topics, subdialogues, robot and child turns and child initiative, per child

The wizard was instructed to first let the child get familiar with the interaction with the robot by doing the Quiz activity. OAT was thus not triggered immediately at the beginning. The wizard was free to decide when to continue pursuing an OAT topic, address a new one or return to the main activity (also upon request from the child). Concerning the OAT topic selection, the wizard was instructed to always raise the issue of the diabetic diary function (but not necessarily as the first theme). When possible, the wizard should prefer OAT topics related to the current semantic content in the ongoing activity (e.g., related to the content of a question in Quiz or to the objects being sorted in SandTray).

5 Off-Activity Talk Analysis

We collected a total of 102 OAT subdialogues. We defined an OAT *subdialogue* as the chunk of interaction around one OAT *topic prompt* by the robot, i.e., the robot raising an OAT topic. An OAT subdialogue ends either by a return to the current activity or by another robot's OAT prompt (typically on a related topic). In other words, we count the robot's OAT prompts. (1) shows a slightly modified excerpt for illustration. It contains two OAT subdialogues: (1a)–(1h) and (1i)–(1n), around the OAT prompts in (1c) and (1k), respectively.

(1)
a. R I'm curious.
b. C Tell me.
c. R Do you check glycemia yourself at home?
d. C Yes, yes,
e. C I'm doing it all by myself.
f. C My parents don't check me, practically never.
g. R Very good.
h. C Thanks.
i. R May I ask more?
j. C Sure. Say all you want.
k. R Do you realize when your blood sugar drops?
l. C Yes, always.
m. C Do you want to tell me anything else?
n. R Not now, maybe later.

OAT subdialogues vary in length, from 2 to 16 turns. There were in total 209 robot and 194 children turns. Tab. 1 shows basic distribution data.

The number of robot and child turns was usually balanced, although there are a few children who were reluctant to respond and show considerably fewer turns than the robot. While most of the children's turns are responses to the robot's OAT inititative, the children sometimes took initiative, too. We will return to this in more detail in §5.3.

5.1 OAT Prompting

The robot raised between 1 and 5 OAT topics per child. The number of OAT subdialogues varied considerably, between 3 and 13.

When preparing the experiment, we expected that the OAT prompt would typically pick up a *contextual topic*, i.e., a topic already present in the context of the current activity. (2) illustrates this: after a nutrition Quiz question involving vegetables in (2a), a nutrition OAT is initiated about eating habits concerning vegetables in (2d).

(2)
a. R Now let's talk about nutrition. What are the best months to eat fresh broccoli?
b. C From June to September.
c. R No, that's wrong. From September to March.
d. R Do you like vegetables?
e. C Not that much.
f. R Which are your favorite foods?
g. C There are a couple - say, pizza.
h. R Yummy.

This was less frequent than we had envisaged (cf. Tab. 2). Among the 56 OAT prompts issued during an activity, only 15 address a contextual topic, whereas 41 address a non-contextual topic.

The remaining 46 prompts are instances of OAT *topic continuation*: an OAT topic is elaborated upon in a sequence of subdialogues. Typically, another facet of a topic is addressed in a continuation. For example, the function of the diary is addressed first, and then the difficulties in filling it in; or a subdialogue about glycemia is followed by one about insulin injections, thus elaborating on the diabetes topic. The OAT prompt (1k) illustrates continuation on the diabetes topic, prompt (2f) a continuation on nutrition.

Relation	Quiz	SandTray	Dance	Total
contextual	5	10	-	15
non contextual	18	21	2	41
continuation	18	25	3	46
Total				102

Table 2: Relation of OAT topic to context

Topic	# no cont.	# cont.	# subdialogues
Hobbies	7	3	3
Diabetes	10	7	11
Nutrition	20	12	14
Friends	2	0	0
Adults	2	0	1
Diary	28	9	17

Table 3: Frequency of topic continuation and number of subdialogues per topic.

Tab. 2 also shows that contextual topics are relatively more frequent in SandTray than in Quiz, and absent in Dance.

Tab. 3 shows how often the addressed OAT topics were continued and the number of subdialogues per topic. The length of single topic chains varies from usually 1 to 3 subdialogues; only in one case the Diabetes topic was elaborated in 4 subdialogues, prompting the subtopics glycemia, insuline injections and injection places.

Tab. 4 shows the frequency of raising the various OAT topics, and also the distribution of OAT topics across the activities. Recall that Quiz was the first activity for each child and that the diary topic should always be raised. It is therefore not surprising that the diary topic is most often raised during Quiz. Quiz is also where the diabetes topic is raised most often. Nutrition, on the other hand, is most often raised in SandTray. This is because questions about food choices and preferences fit well into the context when the child is sorting edible items. That is also why we find more contextual topics here.

OAT was triggered only in very few cases during Dance, mostly raising non-contextual topics. Just in two cases a previous topic was continued: as a child didn't understand a question about the diary during the Quiz game, the topic was raised again during Dance (*I'm curious. We were talking about the food diary. Do you remember to fill it in?*) and again continued in a second subdialogue. In another case, the Dance activity concluded with a Diary reminder.

Although the diary topic was in a sense obligatory, there are only 4 cases where it is raised as the first OAT topic. Hobbies, diabetes and nutrition were the other topics raised first.

Topic	# Subtopics	Quiz	SandTray	Dance
Hobbies	10	0	7	3
Diabetes	17	12	5	0
Nutrition	31	2	28	1
Friends	2	0	2	0
Adults	2	0	2	0
Diary	37	20	13	4
Other	3	0	3	0

Table 4: Frequency and distribution of OAT topics

Topic	Pos.	Neg.	Short	Full	Elab	None
Hobbies	1	-	6	1	2	-
Diabetes	6	-	4	5	2	-
Nutrition	10	4	5	8	2	2
Friends	2	-	-	-	-	-
Adults	1	-	-	1	-	-
Diary	10	2	5	7	9	4
Other	-	1	-	-	-	2

Table 5: Form of children's responses

5.2 Childrens' Responses to OAT Prompts

Tab. 5 shows the distribution of children's responses to OAT prompts. First of all, the children mostly did respond. We shall say more about engagement in §5.4, here we concentrate on the surface form and content of the responses.

Brief responses prevail, including yes/no and their equivalents (cf. (1d)) and short responses (typically phrases), e.g., naming a food. This reflects the fact that OAT prompts were most often formulated as closed questions, allowing such short answers (e.g., (1c), (1k) again). Nevertheless, full-sentence responses such as (1e) are as frequent as short-phrase ones, and have a similar distribution across topics. There is of course variation across children: some gave no full response whereas others gave a few. Moreover, children seem to give more detailed answers during Quiz than during the SandTray activity; maybe because Quiz is actually *interrupted* by the OAT prompt, while SandTray usually goes on in parallel.[2]

On the other hand, the instances where children elaborated on their response, as in (1f) for example, are fewer and not equally distributed: most occurred in response to the general prompt about the diary topic, shown in (3).

(3) R I know Gabriella gave you a food diary to fill in, it is very interesting. Would you explain to me how it works?

[2]In some cases the child's answer is even interrupted by the game-related feedback.

i. C1 Yes. You enter the meals you eat and the blood sugar before and after eating.

ii. C2 No - I don't remember myself so well how to fill it in.

(3i) exemplifies a positive elaborated answer. However, most of the other elaborations on this topic were indications of problems, e.g., the child did not know how to fill in the diary, as in (3ii), or has not yet received it.

5.3 Childrens' OAT Initiative-Taking

We were delighted to observe that the children sometimes grabbed the initiative and raised an OAT topic themselves. It happened in 20 subdialogues, which we consider a high occurrence compared with the 102 robot's OAT subdialogues.

In 17/20 cases the child asked something about the robot. Most frequent (10/17) was *reciprocation* of OAT topics, such as nutrition, e.g., asking whether the robot also likes a given food (3); what it's favorite food is (2); how often it eats fruit (1); or hobbies, e.g., what the robot does in spare time (2); whether it does any sport (1); whether it likes a given sport (1). In the other cases (7/17) the child brought up a *new topic*: what is the robot's name (1)[3]; how old it is (1); does the robot have diabetes (2); why would it need a diary (1); why should it be afraid of needles (1); how does it know how cheese tastes (1). A new topic was sometimes brought up out of the blue and sometimes as the child's follow-up to something the robot said.

In the 3/20 cases where the child's initiative was not asking something about the robot, the child asked whether the robot wanted to discuss any other topic (1); requested an explanation of the diary function (1); and disputed whether pizza was a good food choice (1).

5.4 Engagement

As already said, children mostly did respond to the OAT prompts. Moreover, noone ever turned down an explicit OAT-permission check, such as that in (1a) or (1i). Not surprisingly, there are individual differences between children in terms of how readily they engage in OAT. There are many instances of OAT where the children engage very naturally, respond friendly and elaborate, reciprocate, and even take the initiative, as discussed above. But there are also cases like (4) where there appears to be reluctance, the responses are terse, and the child may request to return to the main activity.

(4) a. R Do you like dancing?
b. C Yes.
c. R Same here.
d. C Shall we play another game?

Only one child very much resisted engagement in OAT altogether. His responses were inaudible, mono-syllabic or simply absent (cf. (5c)–(5f)):

(5) a. R What's the weirdest food you've ever tried?
b. C Hot dog.
c. R Is it good?
d. R What are your favorite foods?
e. R I know Gabriella gave you a food diary to fill in, it is very interesting. Would you explain to me how it works?
f. R If you don't want, doesn't matter. Just remember to fill it in, it's important.
g. C Okay.
h. R Is there something you'd like to ask me?
i. C No.
j. R Okay.

We carried out an annotation of engagement: 3 annotators (native speakers of Italian, 2 of the authors and the psychologist who was involved in the experiment) annotated every OAT subdialogue on a 5-point scale: 1 - passive, uninterested (not participating); 2 - annoyed; 3 - courteus, neutral; 4 - interested; 5 - very interested.

To measure inter-rater agreement, we computed Kendall's *W*. It was quite high at a value of 0.87. The annotators reported that the most important factors for the annotation are intonation and the manner of the answer.

We observed informally a decline in the childrens' engagement in OAT later in the interactions. This was particularly so for OAT about diary or diabetes, while topics like free time or nutrition (favorite foods, weirdest food) were usually answered more willingly. However, there is no statistically confirmed general trend that engagement drops. It is clearly true for some children, while others maintain more or less the same level. Due to large individual variety we cannot conclude what system behavior triggers engagement.

It may be tempting to use the number of turns or subdialogues as a measure of the child's engagement in OAT. However, this is not the case, because sometimes the robot asks more times to get a satisfying answer. All annotators found that the *most positive* interaction is the one in which the child speaks with the robot as if it were a *real*

[3]The robot normally introduces itself by name at the beginning of the interaction. In this case, the session was continued after a technical problem.

play mate and not just a robot. This child has as many turns as others who seemed to become annoyed at the end of the interaction. Full responses do not appear to correlate with engagement either, but rather with the topics and the question type.

6 DISCUSSION AND CONCLUSIONS

In this paper we present the analysis of OAT subdialogues collected in a WOZ experiment with a conversational human-robot interaction system designed to provide, through different activities, useful contents to children with type I diabetes with the aim to help them in managing their condition. We investigate the distribution and character of OAT subdialogues and the responses of the children to the system-triggered OAT stimulation and observe the following: (1) children generally respond to the robot's prompt; (2) majority of full and elaborated responses occured on the diabetes topic; (3) the majority of responses on other topics are brief, which is likely at least partially due to their formulation of the prompts as closed questions; (4) a valuable number of children initiated OAT addressing the robot, thus making obvious the requirement to formulate a consistent background story for the robot character as part of the OAT design; (5) most of the children conducted the dialogue with the robot in a very natural way (e.g., they were engaged and interested, reciprocated OAT); (6) the engagement of some children decreased with the progress of the interaction.

Apparent lack of engagement is hard to interpret, because it is impossible to distinguish between disinterest in OAT topics as such (e.g., due to personality traits), or a reluctance to disclose personal information, or simple interest in and concentration on the main activity. Regarding the observed decreased interest in OAT with the progress of the interaction, we have also to take into account the fact that the system response was often extremely delayed or fragmentary and the synthesized speech output was hard to understand for long/complex utterances. Our aim in the near future is to automate OAT, so as to avoid long waiting times due to the wizard's typing.

On the other hand, the results obtained in this study are admittedly idealized due to the fact that there was no noise due to speech recognition and/or interpretation errors. In future work we need to study strategies for coping with these, as well as possible alternative OAT strategies and the adaptation of the system behavior to that of the child, in various respects.

Besides engagement, OAT has also a tangible effect on the relationship building process: observers (the psychologist and experimenters) note that when the robot asks more personal questions focused on the child, the child becomes curious and surprised. In a number of cases this leads to reciprocal questions, so as to start a "real" conversation with a friend who cares about their interests, habits, feelings, thus corroborating the evidence presented in (Kruijff-Korbayová et al., 2014). The fact that the children ask similar questions suggests that they imagine that the robot can have similar habits and preferences (even also about food or having diabetes, which is irrational if we consider it disengaged from the coversation). This perceived "humanization" of the robot fosters the concept of OAT as a means for observation and eliciting self-disclosure by the care givers, exerting a different approach in a sort of engaging and warming interaction (from an emotional point of view) and triggering, for example, a positive interplay between the establishment of a relationship and the adherence to specific medical guidelines.

In (Kruijff-Korbayová et al., 2014) we report findings concerning the overall effect of OAT: We have qualitative evidence that the presence of OAT during the individual interactions is linked both to a positive effect on the children's perception of the robot, inducing them to see it as a friend and then feeling free and at ease during the playing session, and to a better adherence towards specific medical guidelines like filling in a nutritional diary. Moreover, we found a statistically significant correlation between the presence of OAT in the interaction and the propensity of children to plan and participate in further interaction(s) with the system, in comparison to the non-OAT condition. An interesting topic for future work is to investigate whether any of the OAT characteristics studied in the current paper correlate with the overall effect of OAT.

ACKNOWLEDGMENT

We thank (1) all experiment participants; (2) our collaborators who helped to develop the system, carry out the experiments and process the data; (3) the Center for Pediatric and Adolescent Endocrinology of Ospedale San Raffaele and the Sostegno70 association for diabetic children for their constant and active support of our research.

References

Carole Adam, Lawrence Cavedon, and Lin Padgham. 2010. Hello Emily, how are you today?: personalised dialogue in a toy to engage children. In *Proc. of the 2010 Workshop on Companionable Dialogue Systems*, pages 19–24. Association for Computational Linguistics.

Aldebaran. 2014. Aldebaran website. `http://www.aldebaran-robotics.com/en`.

Aliz-E. 2014. ALIZ-E website. `http://aliz-e.org/`.

Paul Baxter, Rachel Wood, and Tony Belpaeme. 2012. A Touchscreen-Based Sandtray to Facilitate, Mediate and Contextualise Human-Robot Social Interaction. In *HRI, Boston, MA, U.S.A*.

T. Belpaeme, P. Baxter, R. Read, Wood., H. Cuayáhuitl, B. Kiefer, S. Racioppa, Kruijff-Korbayová G. I., Athanasopoulos V. Enescu, R. Looije, M. Neerincx, Y. Demiris, R. RosEspinoza, A. Beck, L. Ca namero, A. Hiolle, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, and R. Humbert. 2013. Multimodal childrobot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):35–53.

T. Bickmore and J. Cassell. 2001. Relational agents: A model and implementation of building user trust. In *Proc. of CHI'01*.

T. Bickmore and T. Giorgino. 2006. Health dialog systems for patients and consumers. *J Biomed Inform*, 39(5):556–571, Oct.

T. W. Bickmore and R. W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293 – 327, Jun.

T. Bickmore, D. Schulman, and L. Yin. 2010. Maintaining engagement in long-term interventions with relational agents. *International Journal of Applied Artificial Intelligence special issue on Intelligent Virtual Agents*, 24(6):648–666.

V. Chidambaram, Vijay, Y.H. Chiang, and B. Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proc. of 17th ACM/IEEE International Conference on. IEEE Human-Robot Interaction (HRI)*.

K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. Koay, and I. Werry. 2005. What is a robot companion - friend, assistant or butler? In *Conference on Intelligent Robots and Systems, Proc. IEEE IRS/RSJ Int.*, pages 1488–1493, Edmonton, Alberta, Canada.

J. Fasola and M. J. Mataric. 2012. Socially assistive robot exercise coach: Motivating older adults to engage in physical exercise. In *In J. P. Desai, G. Dudek, O. Khatib and V. Kumar (eds.), ISER , Springer. ISBN: 978-3-319-00064-0*, pages 463–479.

R. Higashinaka, K. Dohsaka, and H. Isozaki. 2010. Effects of self-disclosure and empathy in human-computer dialogue. In *Proc. of IEEE Spoken Language Technology Workshop, Goa, India*.

Peter H. Kahn, Nathan G. Freier, Takayuki Kanda, Hiroshi Ishiguro, Jolina H. Ruckert, Rachel L. Severson, and Shaun K. Kane. 2008. Design patterns for sociality in human-robot interaction. In *Proc. of the 3rd ACM/IEEE international conference on Human robot interaction (HRI '08).*, pages 97–104.

T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. 2004. Interactive robot as social partners and peer tutors for children:a field trial. In *Human-Computer Interaction*, pages 19:61–84.

Cory D. Kidd and Cynthia Breazeal. 2007. A robotic weight loss coach. In *Proc. of Twenty-Second Conference on Artificial Intelligence (AAAI)*, Vancouver, British Columbia, Canada.

T. Klüwer. 2011. "I like your shirt" – dialogue acts for enabling social talk in conversational agents. In H. et al. Vilhjálmsson, editor, *Proc. of IVA*, pages 14–27, Berlin Heidelberg. Springer-Verlag.

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. 2012a. A conversational system for multi-session child-robot interaction with several games. In *German Conference on Artificial Intelligence (KI)*. system demonstration description.

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. 2012b. Spoken language processing in a conversational system for child-robot interaction. In *Workshop on Child-Computer Interaction*.

I. Kruijff-Korbayová, E. Oleari, I. Baroni, B. Kiefer, M. Coti Zelati, C. Pozzi, and A. Sanna. 2014. Effects of off-activity talk in human-robot interaction with diabetic children. In *Proceedings of the Ro-Man Conference*, Edinburgh, UK.

Kwan Min Lee, Wei Peng, Seung-A. Jin, and Chang Yan. 2006. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, 56(4):754–772.

Y. Moon. 2000. Intimate exchanges: using computers to elicit self-disclosure from consumers. *The Journal of Consumer Research*, 26(4):323–339.

M. Nalin, I. Baroni, A. Sanna, and C. Pozzi. 2012. Robotic companion for diabetic children: emotional and educational support to diabetic children, through an interactive robot. In *Proc. of the 11th International Conference on Interaction Design and Children*, pages 260–263.

H. T. Reis and P. Shaver. 1998. Intimacy as an interpersonal process. In *in Handbook of personal relationships John Wiley and Sons*, pages 367–398.

Raquel Ros, Yiannis Demiris, Ilaria Baroni, and Marco Nalin. 2011. Adapting Robot Behavior to User's Capabilities: a Dance Instruction Study. In *HRI (Human-Robot Interaction) Conference 2011*.

Raquel Ros, Alexandre Coninx, Georgios Patsis, Valentin Enescu, Hichem Sahli, and Yiannis Demiris. 2014. Behavioral accommodation towards a dance robot tutor. In *Proc. of the 9th International Conference on Human-Robot Interaction (HRI)*. To appear.

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.

Sostegno70. 2014. Sostegno70 website. `http://www.sostegno70.org`.

Using subtitles to deal with Out-of-Domain interactions

Daniel Magarreiro, Luísa Coheur, Francisco S. Melo
INESC-ID / Instituto Superior Técnico
Universidade Técnica de Lisboa
Lisbon, Portugal
name.surname@tecnico.ulisboa.pt

Abstract

This paper explores the possibility of using interactions between humans to obtain appropriate responses to Out-of-Domain (OOD) interactions, taking into consideration several measures, including lexical similarities between the given interaction and the responses. We depart from interactions obtained from movie subtitles, which can be seen as sequences of turns uttered between humans, and create a corpus of turns that can be used to answer OOD interactions. Then, we address the problem of choosing an appropriate answer from a set of candidate answers, combining several possible measures, and illustrate the results of our approach in a simple proof-of-concept chatbot that is able deal with OOD interactions. Results show that 61.67% of the answers returned were considered plausible.

1 Introduction

Recent years have witnessed the appearance of *virtual assistants* as a ubiquitous reality. Well-known examples include *Siri*, from Apple, *Anna*, from IKEA, and the buttler *Edgar Smith*, at Monserrate Palace (see Fig. 1).

Such systems are typically designed to interact with human users in well-defined domains, for example by answering questions about a specific subject or performing some pre-determined task. Nevertheless, users often insist in confronting such domain-specialized virtual assistants with OOD inputs.

Although it might be argued that, in light of their assistive nature, such systems should be focused in their domain-specific functions, the fact is that people become more engaged with these applications if OOD requests are addressed (Bickmore and Cassell, 2000; Patel et al., 2006).

Figure 1: The virtual buttler, Edgar Smith, which can be found at Monserrate Palace, in Sintra, Portugal (Fialho et al., 2013).

Current approaches are able to address specific OOD interactions by having the system designer handcraft appropriate answers. However, it is unlikely that system designers will be able to successfully anticipate all the possible OOD requests that can be submitted to such agents. An alternative solution to deal with OOD requests is to explore the (semi-)automatic creation/enrichment of the knowledge base of virtual assistants/chatbots, taking advantage of the vast amount of dialogues available at the web. Examples of such dialogues include those in play/movie scripts, already used in some existing systems (Banchs and Li, 2012).

In this paper, we follow (Ameixa et al., 2014) and adopt an alternative source of dialogues, namely *movie subtitles*. The use of movie subtitles brings two main advantages over scripts and other similar resources. First, the web offers a vast number of repositories with a comprehensive archive of subtitle files. The existence of such collection of subtitle files allows data *redundancy*, which can be of great help when selecting the adequate reply to a given OOD request. Secondly, subtitles are often available in *multiple languages*, potentially

enabling multilingual interactions.[1]

Our approach can be broken down into two main steps, representing our contributions. First, we describe the process of building an improved version of *Subtle*, a corpus of interactions, created from a dataset of movie subtitles. Secondly, we describe a set of techniques that enables the selection/retrieval of an adequate response to a user input from the corpus. The proposed techniques are deployed in a dialogue engine, the *Say Something Smart* (SSS), and an evaluation is conducted illustrating the potential behind the proposed approach in addressing OOD interactions.

This paper is organised as follows. Section 2 surveys some related work. Section 3 describes the construction of the *Subtle* corpus. The SSS engine is described in Section 4 and Section 5 presents the results of a preliminary evaluation. Section 6 concludes, pointing directions for future work.

2 Related work

Virtual assistants have been widely used to animate museums all over the world. Examples include the *3D Hans Christian Andersen* (HCA), which is capable of establishing multi-modal conversations about the namesake writer's life and tales (Bernsen and Dybkjaer, 2005), *Max*, a virtual character employed as guide in the Heinz Nixdorf MuseumsForum (Pfeiffer et al., 2011), the twins *Ada* and *Grace*, virtual guides in the Boston Museum of Science (Traum et al., 2012) and *Edgar Smith* (Fialho et al., 2013), a virtual butler that answers questions about the palace of Monserrate, in Sintra, Portugal (see Fig. 1).

However, and despite the sophisticated technology supporting these (and similar) systems, they are seldom able to properly reply to interactions that fall outside of their domain of "expertise"[2], even though such interactions are reported as quite frequent. For instance, Traum et al. (Traum et al., 2012) report that 20% of the interactions with *Ada* and *Grace* are inappropriate questions.

In order to cope with such OOD interactions, several approaches have been proposed in the literature. For example, when unable to understand a specific utterance (and formulate an adequate answer), *Edgar* (Fialho et al., 2013) suggests questions to the user. In the event that it is repeatedly unable to understand the user, *Edgar* starts talking about the palace. Finally, in order to mitigate the effect of such misunderstandings on the user's engagement and perception of agency, *Edgar* was designed to "blame" his age and bad hearing for its inability to understand the user. In a different approach, *HCA* (Bernsen and Dybkjaer, 2005) changes topic when lost in the conversation. Also, much like *Edgar*, *HCA* has been designed with an "excuse" for not answering some questions: the "virtual HCA" does not yet remember everything that the "real Hans Christian Andersen" once knew. *Max* (Pfeiffer et al., 2011) consults a web-based weather forecast when queried about the weather, and Wikipedia, when approached with factoid questions (Waltinger et al., 2011). In (Henderson et al., 2012), a set of strategies to deal with non understandings is proposed.

Recently, Banchs and Li introduced *IRIS* (Banchs and Li, 2012), a chat-oriented dialogue system that includes in its knowledge sources the *MovieDiC* corpus (Banchs, 2012) . The *MovieDiC* corpus consists of a set of interactions extracted from movie scripts that provides a rich set of interactions from which the system can select a plausible reply to the user's input.

In this paper we take this idea one step further, and propose the use of movie subtitles to build a corpus for open-ended interactions with human users. Subtitles are a resource that is easy to find and that is available in almost every language. In addition, as large amounts of subtitles can be found, linguistic variability can be covered and redundancy can be taken into consideration (if a turn is repeatedly answered in the same way, that answer is probably a plausible answer to that turn).

3 From subtitles to interactions: Building the *Subtle* corpus

In this paper we use knowledge bases constituted of *interactions*, an approach already used in other existing systems (Traum et al., 2012). Each interaction (adjacent pair) comprises two turns, (T, A), where A corresponds to an answer to T, the *trigger*.[3] The following are examples of interactions:

[1] In this paper, we will focus on English, although some experiments with Portuguese were also conducted.

[2] Check `http://alicebot.blogspot.pt/2013/07/turing-test-no-sirie.html` to see *Siri* (Apple's virtual assistant) answers to the 20 questions of the 2013 Loebner Prize contest.

[3] We use the word *trigger*, instead of the usual designation of *question*, since not every turn includes an actual question. Throughout the text, we also use the designations *input* and

```
(T1: You know, I didn't catch your age.
     How old are you?,
 A1: 20)
```

```
(T2: So how old are you?,
 A2: That's none of your business)
```

In this section we describe the process of building interaction pairs based on movie subtitles. We designed a configurable process for building the corpus that takes into consideration the language of the subtitles being processed (henceforth, English and Portuguese) and other elements that should be considered when building the corpus, such as the time elapsed between two consecutive subtitles. Independently of the particular configuration adopted, we refer to the corpus thus built as *Subtle*, although different configurations will evidently lead to different corpora. This corpus is an improved version of the one described in (Ameixa and Coheur, 2014) and (Ameixa et al., 2014).

3.1 Subtitles: The starting point

We obtained 2Gb of subtitles in Portuguese and English from *OpenSubtitles*.[4] These files are in the `srt` format, which consists of a sequence of slots, each containing the following information:

1. The *position* of the slot in the sequence.
2. The *time* indicating when the slot should appear/disappear on the screen.
3. The *content* of the subtitle.

A blank line indicates the start of a new slot. An example of a snippet from a subtitle's file is depicted in Fig. 2.

The 2Gb of subtitle data used includes many duplicate movie subtitles that were removed. In particular, we obtained a total of $29,478$ English subtitle files corresponding to a total of $5,764$ different movies. In removing the duplicate entries, we selected the subtitle file containing the largest number of characters. Similarly, we obtained a total of $14,679$ Portuguese subtitle files corresponding to a total of $3,701$ different movies. In the end, the *Subtle* corpus was built from $5,764$ English subtitle files and $3,701$ Portuguese subtitle files.

request to refer to user turns.

[4] `http://www.opensubtitles.org/`

```
770
01:01:05,537 --> 01:01:08,905
And makes an offer so ridiculous,

771
01:01:09,082 --> 01:01:11,881
the farmer is forced to say yes.

772
01:01:12,752 --> 01:01:15,494
We gonna offer to buy Candyland?
```

Figure 2: Snippet of a subtitle file.

3.2 Extracting interactions from subtitles

We now describe the process of extracting interactions from the selected subtitles files.

Cleaning data

Besides the actual subtitles, there is information provided in the subtitle files that is irrelevant for dialogue and should, therefore, be removed. Examples of portions removed include those containing:

Characters' names. Some subtitle files include the name of the speaker at the beginning of the utterance (e.g., *Johnny: Oh hi, Mark.*). This is particularly useful both when a character is not appearing on the screen and for hearing impaired watchers. Since such names should not be included in the responses of our system, they were eliminated in every turn they appear.

Sound descriptions for hearing impaired. It is also common for subtitle files to include the sound descriptions being played that are relevant for the watcher to perceive (e.g. `[TIRES SCREECHING]`). Such descriptions are not actual responses, so we removed them from the corpus.

Font-changing tags. Subtitles sometimes include tags that video players can interpret to change the normal font in which the tagged subtitle is to be displayed (e.g. `<font color="#ffff00" size=14> Sync by honeybunny </font>`). Such tagged subtitles seldom contained any dialogue element and, therefore, were eliminated.

Finding real turns

The main challenge in building the *Subtle* corpus is to decide which pairs of consecutive slots in the subtitle file correspond to an actual dialogue and which ones do not (and instead correspond, for instance, to a scene change).

In contrast to the version of *Subtle* described in (Ameixa et al., 2014), we allow the user to configure the maximum time allowed between two slots for them to be considered part of a dialogue and used to build an interaction pair. For example, if that time is set to 1 second and two slots are separated by more than that period, they will not be considered as an interaction pair. However, a hard time threshold is difficult to set appropriately, and may lead to useful interactions being discarded from the corpus, if the corresponding value is not adequately set.

To mitigate the impact of a hard time threshold, we also allow the possibility of setting the value of the maximum time between slots to 0, in which case *all* consecutive pairs of slots are considered to be part of a dialogue and used to construct an interaction pair. This latter option ensures that the corpus will contain all the information in the subtitles, but also means that many interaction pairs that are not real interaction pairs in a dialogue will be present in the corpus. As will soon become apparent, we compensate for this disadvantage by including a "soft threshold" mechanism when choosing an answer from a set of possible answers.

Another challenge in processing the subtitles stems from the fact that there is not a standard formatting followed by all the subtitle creators. To handle these formatting differences, we identified common formatting patterns in the process of building the *Subtle* corpus, and specialised, handcrafted rules were designed to take care of such patterns. For instance, when two consecutive subtitle slots correspond to excerpts of a sentence spoken by one single character, the first utterance usually ends with an hyphen, a comma or colon, and the second starts in lowercase.

The snippet of Figure 2 illustrates the aforementioned situation, and a rule has been designed to address it, resulting in the interaction:

```
(T3: And makes an offer so
     ridiculous, the farmer is
     forced to say yes.,
A3: We gonna offer to buy
    Candyland?)
```

We refer to (Ameixa and Coheur, 2014) for additional details on other rules.

Finally, we note that the context of each turn is kept while building of the *Subtle* corpus. Although such context information is currently not used in the dialogue system described ahead, it is still kept as it may provide useful information for future improvements of the dialogue system. An excerpt of the resulting *Subtle* corpus is provided in Fig. 3.

```
SubId - 100000
DialogId - 1
Diff - 3715
T - What a son!
A - How about my mother?

SubId - 100000
DialogId - 2
Diff - 80
T - How about my mother?
A - Tell me, did my mother
      fight you?

SubId - 100000
DialogId - 3
Diff - 1678
T - Tell me, did my mother
    fight you?
A - Did she fight me?
```

Figure 3: Excerpt of the *Subtle* corpus obtained from the subtitle files.

In the example depicted in Fig. 3, `SubId` is a number that uniquely identifies the subtitle file from which the corresponding interaction was extracted. `DialogId` is a value used to find back-references to other interactions in the same conversation (the context). `Diff` is the difference in time (in milliseconds) between the trigger and the answer as registered in the subtitle file. Finally, T and A are the trigger and the answer, respectively. Note that, in the second interaction featured in the example of Fig. 3, it is very likely that both the trigger and the answer are spoken by the same character. This observation is also supported by the fact that the time difference between trigger and answer is very small. As already mentioned, the time difference will be taken into consideration when selecting the answer to an input by the user, both by weighting down answers with a time

difference that is too small (as in the example) or too large.

3.3 The Subtle Corpus: Some numbers

Table 1 summarizes some information regarding the *Subtle* corpus, generated when the time threshold between two slots is set to 0.

Table 1: Summarized information regarding the *Subtle* Corpus.

English

# Movies	# Movies ok	# Interactions	Average
5,764	5,665	5,693,811	1,005

Portuguese

# Movies	# Movies ok	# Interactions	Average
3,701	3,598	3,322,683	923

Some subtitle files did not comply with the usual `srt` format and were discarded. In English, from the initial 5,764 subtitle files (listed under **# Movies** in Table 1), 99 were discarded and only 5,665 files were used (listed under **# Movies ok** in Table 1). In Portuguese, from the initial 3,701 files, 3,598 were used to build the corpus. The processing of these files resulted in a total of 5,693,811 English interaction pairs (listed under **# Interactions** in Table 1) and 3,322,683 Portuguese interaction pairs, with an average number of interactions per file of 1,005 for English and 923 for Portuguese (**# Average** in Table 1).

4 The *Say Something Smart* Engine

In this section we describe the process of choosing an answer, being given an input from the user. When a user poses his/her request, this input is matched against the interactions in the *Subtle* corpus, and a set of answer candidates is retrieved. Then, a response needs to be chosen from the candidate answers. To this end, we index the *Subtle* corpus and extract a set of candidates; we score these candidates considering several measures and finally return the answer corresponding to the one attaining the best score.

In the continuation, we describe the indexing and selection process in further detail.

4.1 Corpora indexing and candidate extraction

Say something smart (SSS) uses Lucene[5] to index the *Subtle* corpus and its retrieval engine to obtain the first set of possible answers, given a user input (Ameixa et al., 2014). Lucene works with tokenizers, stemmers, and stop-word filters. We used the default ones for English, and the snowball analyzer for the Portuguese language.[6]

In the following we illustrate some of the retrieved interactions, considering the user input "*Do you have a brother?*":

```
(T4: You don't have to go,
       brother.,
 A4: I'm not your brother.)

(T5: You have a brother?,
 A5: Yeah, I've got a brother,
      man. You know that.)

(T6: Joe doesn't have a brother?,
 A6: No brother.)

(T7: Brother, do you have tooth
     paste?,
 A7: What brother?)

(T8: Have you seen my brother?,
 A8: He's not your brother
    anymore.)
```

The example above illustrates one of the problems of choosing an appropriate answer. As it can be seen, many of the interactions returned by Lucene have triggers that are not really related with the given input.

4.2 Choosing the answer

Given a user request u, Lucene retrieves from the set I of all interactions a subset U of N interactions, $U = \{(T_i, A_i), i = 1, \ldots, N\}$. Each interaction $(T_i, A_i) \in U$ is scored according to each of a total of four measures. The final score of each answer A_i to the user input u, $score(A_i, u)$, is computed as a weighted combination of the 4 scores $M_j, j = 1, \ldots, 4$:

$$score(A_i, u) = \sum_{j=1}^{4} w_j M_j(U, T_i, A_i, u), \quad (1)$$

where w_j is the weight assigned to measure M_j.[7]

The four measures implemented are described in the following.

[5] `http://lucene.apache.org`

[6] `http://snowball.tartarus.org/`

[7] All the measures to be applied and the associated weights can be defined by the user.

Trigger similarity with input The first measure, M_1, accounts for the *Jaccard similarity* (Jaccard, 1912) between the user input and the trigger of the interaction. For instance, given the input "*What's your mother's name?*", and the interactions:

```
(T9: How nice. What's your
     mother's name?,
 A9: Vickie.)

(T10: What was your
     mother's name?,
 A10: The mother's name
   isn't important.)
```

M_1 will assign a larger value to the second interaction, since "*What's your mother's name?*" is more similar to T10 than to T9, according with the Jaccard measure.

The measure M_1 is particularly important since, as previously discussed, many of the interactions returned by Lucene have triggers that have little in common with the given input. For example, and considering once again the previous input ("*What's your mother's name?*") some of the triggers retrieved by Lucene were:

```
T11: What's your name?

T12: What's the name your mother
   and father gave you?

T13: Your mother? how dare
   you to call my mother's name?.
```

Response frequency The second measure, M_2, targets the response frequency, giving a *higher score to the most frequent answer*. That is to say, we take into consideration the corpus redundancy. We do not force an exact match and the Jaccard measure is once again used to calculate the similarity between each pair of possible answers. Consider, for example, the request "*How are you?*" and the interactions:

```
T14: Where do you live?
A14: Right here.

T15: Where are you living?
A15: Right here.

T16: Where do you live?
A16: New York City.

T17: Where do you live?
A17: Dune Road.
```

M_2 will give more weight to the answer *Right here*, as it is more frequent than the others.

Answer similarity with input We also take into consideration the answer similarity (Jaccard) to the user input. Thus, M_3 computes the similarity between the user input and each of the candidate answers (after stop words removal). If scores are higher than a threshold it is considered that the answer shares too much words with the user input, and a score 0 is given to the answer; otherwise, the attained similarity result is used in the score calculus, after some normalisations.

Time difference between trigger and answer Finally, we can use in this process the time difference between the trigger and the answer (measure M_4). If there is too much time elapsed between the trigger and the answer, it is possible that they are not a real interaction.

$\diamondsuit$

To conclude, we refer that in (Ameixa et al., 2014) a hard-threshold is used to filter the interactions returned by Lucene considering a similarity measure; the most similar answer is used to decide which response is returned, much like our measure M_2. In this paper, we do not apply any hard-threshold. Instead, we combine a set of four different measures to score the candidates and select the one attaining the largets combined score.

5 Evaluation

In this section we describe some preliminary experiments conducted to validate the proposed approach.

5.1 Evaluation setup

Filipe, depicted in Fig. 4, is a chatbot previously built to allow user interactions with the SSS engine (Ameixa et al., 2014). It is on-line since November 2013.[8]

Using Filipe, we have collected a total of 103 requests made to the original SSS engine by several anonymous users. From this set, we removed

[8]It can be tested in `http://www.l2f.inesc-id.pt/~pfialho/sss/`

Figure 4: Filipe, a chatbot based on SSS.

the duplicates and randomly selected 20 inputs as a test set for our system.

5.2 Are subtitles adequate?

We started our evaluation with a preliminary inspection of *Subtle*, in order to understand if adequate responses could be found there. Three human annotators evaluated the first 25 answers returned by Lucene to each one of the 20 requests from the test set. For each request the annotators would indicate whether *at least one appropriate answer* could be found in these 25 candidate answers returned by Lucene.

The first annotator considered that 19 of the user requests could be successfully answered and that one could not, corresponding to the input "*What country do you live?*".

The second annotator agreed with the first annotator in 19 of the test cases. The only different test case corresponded to the input "*Are you a loser?*", to which the second annotator determined no suitable answer could be found in the ones returned by Lucene.

The third annotator disagreed with both annotators one and two with respect to the input "*What country do you live?*", as he considered "*It depends.*" to be a plausible answer. Additionally, this annotator considered that there was no plausible answer to the input "*Where is the capital of japan?*", to which the other two annotators agreed that "*58% don't know.*" was a plausible answer. Finally, the first and third annotators agreed that "*So what? You want to hit me?*", "*Your thoughtless words have made an incredible mess!*" or "*Shut up.*" would be appropriate answers to "*Are you a loser?*".

Despite the lack of consensus in these test cases, the fact is that the three annotators agreed that 17 out of 20 turns had a plausible answer in the set of answers retrieved by Lucene from the *Subtle* corpus, which is an encouraging result.

The next step is then to study the best way to select a plausible answer from the set of candidate answers retrieved by Lucene. Our framework, presented in Section 4, is evaluated in the continuation.

5.3 Answer selection

We tested five different settings $(S_1, \ldots, S_5)$ to score each interaction pair:

- S_1 – Only takes into account M_1.
- S_2 – Only takes into account M_2.
- S_3 – Takes into account M_1 and M_2.
- S_4 – Takes into account M_1, M_2 and M_3.
- S_5 – Takes into account all four measures.

For the settings S_{1-4} all measures considered were given the same weight. For S_5, however, the weights were optimized experimentally, yielding:

- 40% weight for M_1.
- 30% weight for M_2.
- 20% weight for M_3.
- 10% weight for M_4.

The test set described in Section 5.1 was again used, and SSS was tested in each of the five settings $S_1, \ldots, S_5$ described above. The best scored answer of each log was returned.

In order to evaluate how plausible the returned answers were, a questionnaire was built. It contained the 20 user request from the test set and the answers given considering each of the settings (duplicate answers were removed). We told the evaluators that those were the requests posed by humans to a virtual agent and the possible answers. They should decide, for each answer, if it made sense or not. Figure 5 shows an extract of the questionnaire. 21 persons filled the questionnaire. Results are summarized in Table 2.

	S_1	S_2	S_3	S_4	S_5
%	39.29	45.24	46.90	61.67	51.19

Table 2: Percentage of plausible answers in each setting.

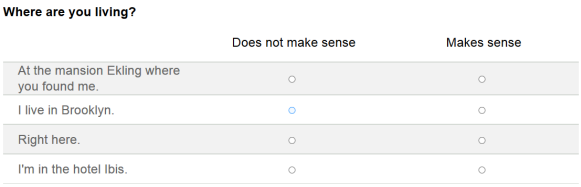


Figure 5: Example of a question in the questionnaire.

We can see that the S_2 setting achieved better results than S_1, and that S_3 (the combination of measures M_1 and M_2) achieved slightly better results than the previous two. This suggests that the combination of the two strategies may yield better results than any of them alone. Moreover S_4 (which added the third measure M_3) achieved the best results, with a difference of almost 15% compared to the strategy of S_3. The last setting (which added the M_4 measure), however, achieved worse results than S_3.

To conclude, our preliminary evaluation suggests that the similarity between the user request and the trigger and the answer should be considered in this process, as well as the redundancy of the answers.

6 Conclusions and future work

As it is impossible to handcraft responses to all the possible OOD turns that can be posed by humans to virtual conversational agents, we hypothesise that conversations between humans can provide some plausible answers to these turns.

In this paper we focus on movies subtitles and we describe the process of building an improved version of the *Subtle* corpus, composed of pairs of interactions from movies subtitles. A preliminary evaluation shows that that the *Subtle* corpus does include plausible answers. The main challenge is to retrieve them. Thus, we have tested several measures in SSS, a platform that, given a user input, returns a response to it. These measures take into consideration the similarities between the user input and the trigger/answer of each retrieved interaction, as well as the frequency of each answer. Also, the time elapsed between the subtitles is taken into consideration. Different weights were given to the different measures and the best results were attained with a combination of the measures: 21 users considered that 61.67% of the answers returned by SSS were plausible; the time elapsed between the turns did not help in the process.

There is still much room from improvement. First, the context of the conversation should be taken into consideration. An automatic way of combining the different measures should also be considered, for instance using a reinforcement learning approach or even a statistical classifier to automatically estimate the weights to be given to each measure. Moreover, semantic information, such as the one presented in synonyms, could be used in the similarity measure; information regarding dialogue acts could also be used in this process.

Also, responses that refer to idiosyncratic aspects of the movie should receive a lower score. Although M_2 can be seen as an indirect metric for this domain-independence (a frequent response is less likely to come with a strong contextual background), responses that include names of particular persons, places or objects should be identified. However, this strategy is not straightforward, as some particular responses containing named entities should not be discarded. This is the case not only to address factoid questions, but also for inputs such as "*Where do you live?*" or "*What is your mother's name?*" whenever a pre-defined answer was not prepared in advance.

Currently we are collecting characters' language models, and intend to use these during the answer candidate selection. Additionally, we are in the process of combining information from movie scripts to enrich subtitles, for example by adding in character names. This added information would enable an easier identification of the dialogue lines of each character as well as creating specific language models; finally, this could also allow us to filter some interaction pairs that represent two lines from the same character.

Acknowledgements

We would like to acknowledge the administrator of OpenSubtitles for providing the subtitle data used in this paper and the anonymous reviewers for helpful comments. This work was partially supported by EU-IST FP7 project SpeDial under contract 611396 and by national funds through FCT – Fundação para a Ciência e a Tecnologia, under projects PEst-OE/EEI/LA0021/2013 and CMUP-ERI/HCI/0051/2013.

References

David Ameixa and Luísa Coheur. 2014. From subtitles to human interactions: introducing the subtle corpus. Technical report, INESC-ID.

David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. Luke, i am your father: dealing with out-of-domain requests by using movies subtitles. In *Proceedings of the 14th International Conference on Intelligent Virtual Agents (IVA'14)*, LNCS/LNAI, Berlin, Heidelberg. Springer-Verlag.

Rafael Banchs and Haizhou Li. 2012. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proc. 50th Annual Meeting of the ACL: System Demonstrations 50th Meeting ACL (System Demonstrations)*, pages 37–42.

Rafael E. Banchs. 2012. Movie-DiC: a movie dialogue corpus for research and development. In *Proc. 50th Annual Meeting of the ACL: System Demonstrations 50th Meeting ACL (Short Papers)*, pages 203–207.

Niels Ole Bernsen and Laila Dybkjaer. 2005. Meet Hans Christian Anderson. In *Proc. 6th SIGdial Workshop on Discourse and Dialogue*, pages 237–241.

Timothy Bickmore and Justine Cassell. 2000. How about this weather? social dialogue with embodied conversational agents. In *Socially Intelligent Agents: The Human in the Loop*, pages 4–8.

Pedro Fialho, Luísa Coheur, Sérgio Curto, Pedro Cláudio, Ângela Costa, Alberto Abad, Hugo Meinedo, and Isabel Trancoso. 2013. Meet Edgar, a tutoring agent at Monserrate. In *Proc. 51st Annual Meeting of the ACL: System Demonstrations*, pages 61–66, August.

Matthew Henderson, Colin Matheson, and Jon Oberlander. 2012. Recovering from Non-Understanding Errors in a Conversational Dialogue System. In *Workshop on the Semantics and Pragmatics of Dialogue*.

P. Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Ronakkumar Patel, Anton Leuski, and David Traum. 2006. Dealing with out of domain questions in virtual characters. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, IVA'06, pages 121–131, Berlin, Heidelberg. Springer-Verlag.

Thies Pfeiffer, Christian Liguda, Ipke Wachsmuth, and Stefan Stein. 2011. Living with a virtual agent: Seven years with an embodied conversational agent at the Heinz Nixdorf MuseumsForum. In *Proc. Re-Thinking Technology in Museums 2011*, pages 121–131.

David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. 2012. Ada and Grace: Direct interaction with museum visitors. In *Proc. 12th Int. Conf. Intelligent Virtual Agents*, pages 245–251.

Ulli Waltinger, Alexa Breuing, and Ipke Wachsmuth. 2011. Interfacing virtual agents with collaborative knowledge: Open domain question answering using Wikipedia-based topic models. In *Proc. 22nd Int. Joint Conf. Artificial Intelligence*, pages 1896–1902.

Generating and Resolving Vague Color References

Timothy Meo[1] and **Brian McMahan**[2] and **Matthew Stone**[2,3]
[1]Linguistics, [2]Computer Science, [3]Center for Cognitive Science
Rutgers University
[1]New Brunswick, NJ 08901-1184, [2]Piscataway, NJ 08854-8019
firstname.lastname@rutgers.edu

Abstract

We describe a method for distinguishing colors in context using English color terms. Our approach uses linguistic theories of vagueness to build a cognitive model via Bayesian rational analysis. In particular, we formalize the likelihood that a speaker would use a color term to describe one color but not another as a function of the background frequency of the color term, along with the likelihood of selecting standards in context that fit one color and not the other. Our approach exhibits the qualitative flexibility of human color judgments and reaches ceiling performance on a small evaluation corpus.

1 Introduction

A range of research across cognitive science, summarized in Section 2, suggests that people negotiate meanings interactively to draw useful distinctions in context. This ability depends on using words creatively, interpreting them flexibly, and tracking the effects of utterances on the evolving context of the conversation. We adopt a computational approach to these fundamental skills. Our goal is to quantify them, scale them up, and evaluate their possible contribution to coordination of meaning in practical dialogue systems.

Our work extends three traditions in computational linguistics. Our approach to semantic representation builds on previous research that emphasizes the context dependence and interactive dynamics of meaning (Barker, 2002; Larsson, 2013; Ludlow, 2014). Our approach to pragmatic reasoning builds on work on referring expressions and its characterization of the problem solving involved in using vague language to identify entities uniquely in context (Kyburg and Morreau, 2000; van Deemter, 2006). Finally, we take a perceptually grounded approach to meaning, which allows us to use empirical methods to induce semantic representations on a wide scale from multimodal corpus data (Roy and Reiter, 2005; Steels and Belpaeme, 2005; McMahan and Stone, 2014).

We present our ideas through a case study of the color vocabulary of English. In particular, we study the problem solving involved in using color descriptors creatively to distinguish one color swatch from another, similar color. In our model, these descriptions inevitably refine the interpretation of language in context. We assume that speakers make choices to fulfill their communicative goals while reproducing common patterns of description. Using corpus data, we are able to quantify how representative of typical English speakers' behavior a particular context-dependent semantic interpretation is.

Our model naturally exhibits many of the preferences of previous work on vague descriptions. For example, the system avoids placing thresholds in small gaps (van Deemter, 2006), that is, in regions of conceptual space that account for little of the probability mass of possible interpretations. In such circumstances, the system prefers more specific vocabulary, where interlocutors are more likely to draw fine distinctions (Baumgaertner et al., 2012). Our approach realizes these effects by simple and uniform decision making that extends to multidimensional spaces and arbitrary collections of vocabulary.

We begin the paper by describing the semantic representation of vagueness in dialogue. Vagueness, we assume, is uncertainty about where to set the threshold in context for the concept evoked by a term. Speakers have the option to triangulate more precise thresholds by interactive strategies such as accommodation, and this helps explain how vague descriptions can be used to refer to objects precisely (van Deemter, 2006).

In Section 3, we describe our model of speakers' decisions in conversation. We focus on speak-

ers that aim to distinguish one thing from another; in these cases, we assume speakers aim to choose a term that's interpreted so that it fits the target and excludes the distractor, while matching broader patterns of language use.

We show how to combine the ideas in Section 4. We formalize the likelihood that a speaker would use a color term to describe one color but not another as a function of the likelihood of selecting standards to justify its application in this context, along with the background frequency of the color term. We describe an implementation of the formalism and report its the qualitative and quantitative behavior in Section 5. It works with a generic lexicon of more than 800 color terms and reaches ceiling performance in interpreting user color descriptions in the data set of Baumgaertner et al. (2012). While substantial additional research is required to explore the dynamics of vagueness in conversation, our results already suggest new ways to apply generic models of the use of vague language in support of sophisticated, open-ended construction of meaning in situated dialogue.

2 The Linguistics of Vagueness

Figure 1 shows an image from a public data set developed to study how people label images with captions (Young et al., 2014). One user chose to distinguish the dogs by calling one brown and the other tan. Another distinguished the dogs by calling one tan and the other white. Each used *the tan dog* to refer to a different dog—yet the way each described the other dog left no doubt about the correct interpretation. This variability and context dependence is characteristic of vagueness in language. The dogs in Figure 1 are borderline cases; there's no clear answer about whether they are tan or not, and speakers are free to talk of either, both, or neither of them as tan, depending on their purposes in the conversation.

In this paper, we explore the descriptive variability seen in Figure 1. How is it that speakers can settle borderline cases in useful ways to move a dialogue forward, and how can hearers recognize those decisions? We won't consider the interactive strategies that interlocutors can use to confirm, negotiate or contest potentially problematic descriptions, although that's obviously crucial for successful reference (Clark and Wilkes-Gibbs, 1986), for coordinated meaning (Steels and Belpaeme, 2005), and perhaps even for meaning itself (Ludlow, 2014). And we won't consider the way multiple descriptions constrain one another, as in Figure 1, although we expect to explain it as a side-effect of holistic interpretive processing (Stone and Webber, 1998). We see our work as a prerequisite for the model building and data collection required to address such issues.

Figure 1: A brown dog and a tan one—or a tan dog and a white one (Young et al., 2014).

In our view, the users of Young et al. (2014) are using *tan* to name color categories. Colors are visual sensations that vary continuously across a space of possibilities. Color categories are classifiers that group regions in color space together (Gärdenfors, 2000; Larsson, 2013). Color terms in English also have another sense, not at issue in this paper, where they refer to an underlying property that correlates with color, as in *red pen* (writes in red ink) (Kennedy and McNally, 2010).

Empirically, color categories seem to be convex regions (Gärdenfors, 2000; Jäger, 2010)—in fact, we model them as rectangular box-shaped regions in hue–saturation–value (HSV) space. Thus, color categories involve boundaries, thresholds or standards that delimit the regions in color space where they apply; context sensitivity can be modeled as variability in the location of these boundaries (Kennedy, 2007). For example, when we categorize the lighter dog of Figure 1 as being distinctive in its color, we must have a color category that fits this dog but not the darker one. This category will group together colors with a suitable interval of yellow hues, suitable low levels of saturation, and suitably high values on the white–black continuum. We can think of this category as one possible interpretation for the word *tan*. By contrast, categorizing the darker dog of Figure 1 as distinctively *tan* involves choosing a category with dif-

ferent thresholds for hue, saturation and value—thresholds that fit the color of the darker dog but exclude that of the lighter one.

When interlocutors use vague terms in conversation, they constrain the way others can use those terms in the future (Lewis, 1979; Kyburg and Morreau, 2000; Barker, 2002). For example, if we hear one or the other dog of Figure 1 described as *tan*, it constrains how we will interpret subsequent uses of the word *tan*. Concretely, we might update the perceptual classifier we associate with *tan* in this context so that it fits the target dog and excludes its alternative (Larsson, 2013). We see this as a case of accommodation, in the sense of (Lewis, 1979).

As speakers, we often count on our interlocutors to accommodate us (Thomason, 1990). We can use vague terms confidently as long as the distinction we aim to draw with them is clear in context and as long as our choice is sufficiently in line with the normal variation in the use of the word, and therefore uncontroversial (Thomason, 1990; van Deemter, 2006). Such criteria seem to support the speaker's choice in Figure 1 to describe either dog as *tan*—provided the speaker provides a complementary description of the other dog. At the same time, if we use language in very unusual ways, we can expect that our interlocutor may have difficulty understanding and may be reluctant to accommodate us. In other words, to use vague language effectively, speakers must be sensitive to whether their utterances update the dialogue context in a natural way.

A common idea in linguistics and philosophy is that knowledge of language associates terms with a probability distribution over categories. This distribution characterizes speakers' information about the likelihood of different possible interpretations for the term that could make sense in context (Williamson, 1996; Barker, 2002; Lassiter, 2009). In other words, vagueness amounts to uncertainty about where to draw boundaries to settle borderline cases.

Thus, when we need to settle borderline cases to generate or understand utterances like *the tan dog*, knowledge of meaning lets us quantify how likely the different resolutions are. In Figure 1, for example, knowledge of language says that *tan* can be interpreted, with a suitable probability, through categories that pick out just the lighter dog, but that *tan* can also be interpreted, with a suitable probability, through categories that pick out just the darker dog. The next section explains how to formalize the reasoning involved in assessing these probabilities, reviews one instantiation of this reasoning for learning semantics, and develops another instantiation for distinguishing colors in context.

3 Rational Analysis of Descriptions

Speakers can use language for a variety of purposes. Their decisions of what to say thus depend on knowledge of language, their communicative situation, and their communicative goals. Following Anderson (1991), rational analysis invites us to explain an agent's action as a good way to advance the agent's goals given the agent's information. When applied to communication, this approach allows us not only to derive utterances for systems but also to infer linguistic representations from utterances when we know the agent's communicative situation and communicative goals.

We apply this methodology to color descriptions in McMahan and Stone (2014). We infer linguistic representations from Randall Munroe's color corpus[1] by assuming that subjects' goals were to say true things and match a target distribution of utterances. These results are available as our Lexicon of Uncertain Color Standards (LUX). We describe this experiment in Section 3.1. We continue in Section 3.2 by creating a new model of the task of creating a distinguishing description. Here the goal is to describe one color, exclude another, and match a target distribution.

3.1 Lexicon of Uncertain Color Standards

Munroe's corpus was gathered by presenting subjects with a color patch and allowing them to freely describe it. It's not interactive language use, but we use it just to model knowledge of meaning. Like all crowdsourced data, Munroe's methodology sacrifices control over presentation of stimuli and curation of subjects' responses for sheer scale of data collection. We work with a subset of data involving 829 color terms elicited over 2.18M trials. Each description is paired with the multiset of color values on which subjects used it. We model the data in HSV space, because color categories generally differ in the *Hue* dimension.

LUX links color descriptions with context-sensitive regions in HSV color space. An example is shown in Figure 2 for the *Hue* dimension.

[1] blog.xkcd.com/2010/05/03/color-survey-results/

The plot shows a scaled histogram of subjects' responses. There is a region on the *Hue* dimension which subjects frequently described as *yellowish green* with borderline cases on either side of it.

To capture the patterns of human responses, the rational analysis approach directly models the uncertainty described in Section 2. For each color term, speakers have possible standards which can be used to partition color space; they are unsure which are at work at any point. For example, the term *yellowish green* only fits those *Hue* values which are above a minimum threshold, $\tau^{Lower,H}$ (or $\tau^{L,H}$ for short), and below a maximum threshold, $\tau^{Upper,H}$ (or $\tau^{U,H}$ for short). We estimate the distribution of possible thresholds; they are shown as the solid black lines in Figure 2.

In choosing to use the color description to fit a point x in HSV space, speakers make a semantic judgment which constrains the possible standards. The naturalness of this judgment is measured in part by the probability mass of possible standards which allow the description to be used. For example, the applicability of *yellowish green* is the probability of the color value x being between the minimum and maximum thresholds in each dimension. For a color description k, this is mathematically defined fully in Equation 1 and more compactly in Equation 2.

$$P(\tau_k^{Lower,H} < x^H < \tau_k^{Upper,H}) \times$$
$$P(\tau_k^{Lower,S} < x^S < \tau_k^{Upper,S}) \times$$
$$P(\tau_k^{Lower,V} < x^V < \tau_k^{Upper,V}) \quad (1)$$
$$= \prod_{d \in \{H,S,V\}} P(\tau_k^{L,d} < x^d < \tau_k^{U,d}) \quad (2)$$

The other factor in subjects' choices is the saliency of the color term. The saliency of color

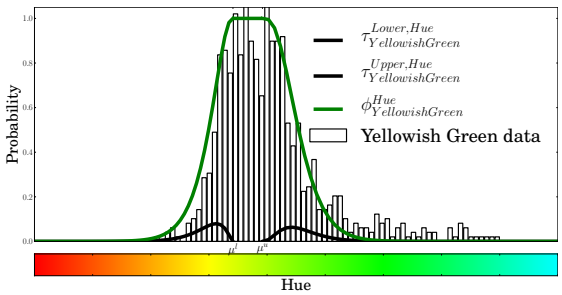


Figure 2: The LUX model for "yellowish green" on the *Hue* axis plotted against a scaled histogram of responses. The ϕ curve, the likelihood of a color counting as "yellowish green", is derived from the τ curves representing possible boundaries.

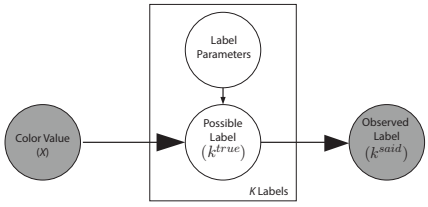


Figure 3: A Bayes Rational Observer sees a color patch. The subjective likelihood $P(k^{true}(c)|c = x)$ describes the likelihood that descriptor k is true of the current color c given that it is located *HSV* point x. The descriptor k is actually said proportional to this subjective likelihood and a weight representing how often a label is said when it is true: $P(k^{said}|k^{true}(c))$. In Munroe's data, the shaded nodes are observed.

description k, also called *availability* and written as $\alpha(k)$, is a background measure of how often the term is used when it is true. Thus, to pick a term that fits a color swatch and use language in a natural way, subjects can pick a color term according to the product of availability and subjective likelihood. Figure 3 summarizes this process in a graphical model.

In Equation 3 , we introduce a simpler notation for Equation 2 that we build on in what follows. We abbreviate $P(\tau_k^{L,d} < x^d < \tau_k^{U,d})$ as $\phi_k^d(x^d)$ and show how $\phi_k^d(x^d)$ can be defined by cases as a function of how x^d is situated with respect to the lower limit $\mu^{L,d}$ and upper limit $\mu^{U,d}$ of the threshold distributions:

$$\phi_k^d(x^d) = \begin{cases} P(x^d > \tau_k^{L,d}), & x^d \leq \mu_k^{L,d} \\ P(x^d < \tau_k^{U,d}), & x^d \geq \mu_k^{U,d} \\ 1, & otherwise \end{cases} \quad (3)$$

LUX was learned from Munroe's data by fitting the parameters of the ϕ function for each description on each dimension independently to the frequency histogram. For example, the parameters for the ϕ function for *yellowish green* in Figure 2 were fit by maximizing the probability that the bins in the data histogram were sampled from the ϕ curve with standard Gaussian noise.

3.2 Distinguishing Descriptions

Munroe's elicitation task is simple; in other settings, people have more complex communicative goals, such as unique reference. These goals modulate the link between internal semantic representation and observed speaker choice. In Munroe's

task, we assume, the speaker sampled from possible descriptive terms based on terms' availability and how likely terms were to fit the target color value. We now consider how this changes when speakers aim to differentiate between two objects.

The literature offers a key insight to get us started: referential expressions are marked as such, and the scalar structure of vague meanings gives strong constraints on how vague terms can be interpreted. For example, *the fat pig* can only refer to the fatter of two pigs in the context, a calculation that is easy to add to algorithms for referring expression generation (Kyburg and Morreau, 2000; van Deemter, 2006). However, things become substantially more complicated in the case of color, because color is multidimensional and color categories can be approximated in competing ways, as with *tan* in Figure 1.

We approach the problem probabilistically. To generate likely unique references, the speaker must sample from possible descriptive terms proportional to terms' availability, how likely terms are to fit the target, and how likely terms are to exclude a distractor. This involves integrating over all possible thresholds, to measure the probability that a description should be interpreted to include one color and exclude another. In the ordinary case where two colors are far enough apart, most thresholds work, and the approach defaults to the kinds of natural descriptions seen in descriptions of colors on their own. However, when the colors become increasingly close, general color descriptions (such as *green*) no longer are likely to signal the distinction we need, while more specific color descriptions are (such as *lime green* and *pale green*). This qualitative behavior is an important part of vague language, as observed by Baumgaertner et al. (2012). (They also suggest that accurate models of color vagueness would be necessary for good performance in difficult cases.)

The same model can inform the resolution of vague descriptions as well as generation. Resolving reference requires reasoning about how well each description applies to each of the candidate referents. We explore this reasoning for generation and understanding in the next section.

4 Algorithm and Implementation

The heart of our method is a measure of the confidence with which we can use a color term to describe a color Y and to exclude a second color Z. We will call this number the Y–but–not–Z confidence rating. This is the probability that the thresholds in context are chosen in such a way that color term k fits color Y but does not fit distractor Z. (That's $P(k^{true}(c)|c = Y) \times P(\neg k^{true}(c)|c = Z)$ in the notation of Figure 3.) To generate a term in context, we might consider each possible color label, calculate its Y–but–not–Z confidence, and finally pick a term proportional to its confidence.

We motivate our mathematical model by considering a single perceptual dimension, most easily visualized as *Hue*. In this case, the Y–but–not–Z confidence is equal to the probability that the upper and lower thresholds of that term can be set such that Y falls inside them, and Z falls outside of them. Thus each confidence rating will involve the multiplication of two values: the probabilities associated with the upper and lower boundaries.

In Figure 4, Y and Z are borderline *Hue* values; both are greener than the typical yellowish green. In this case, there's no constraint on the lower threshold; the lower threshold fits the description with probability 1. On the other hand, only the upper shaded region of Figure 4 supports a categorization of Y but not Z as yellowish green. This area is equal to $\phi(Y) - \phi(Z)$. This is the probability that the *Hue* boundaries for this color term will include Y and exclude Z. Symmetrical reasoning applies in the mirror-image case when the colors are borderline yellow.

Another case is shown in Figure 5, in which Y and Z fall on opposite sides: Y is borderline green, while Z is borderline yellow. In these *contrasting borderline* cases, it's up to the speaker whether to count Y in and Z out or vice versa, as in Figure 1. The choices can be good or bad, however,

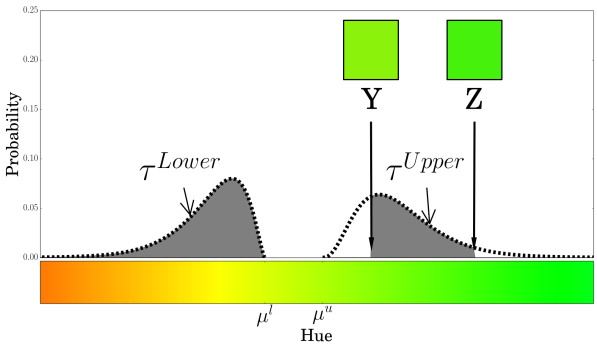


Figure 4: The thresholds that separate two nearby borderline cases cover probability $\phi(Y) - \phi(Z)$, here $0.74 - 0.05 = 0.69$.

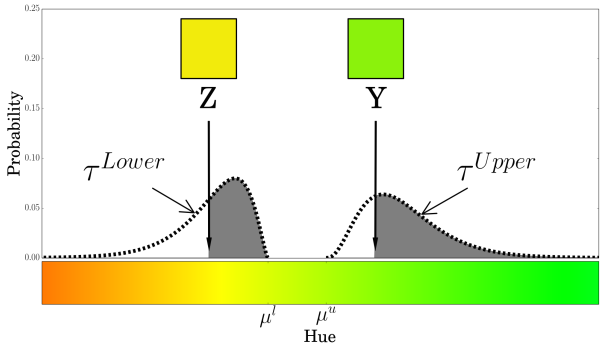


Figure 5: The thresholds that separate two contrasting borderline cases cover probability $\phi(Y) * (1-\phi(Z))$, here $0.74 * (1-0.38) = 0.46$.

because they constrain the context. The probability that the upper threshold includes Y is $\phi(Y)$. The shaded area above Z represents the probability that the lower threshold is placed such that Z is excluded; its area is equal to $1-\phi(Z)$. Thus, the Y–but–not–Z confidence rating for this case is $\phi(Y) * (1-\phi(Z))$. Again, there is a symmetrical case with the colors reversed.

Finally, if Y is not a borderline case, as in Figure 6 then Y does not constrain the thresholds at all. Thus, the Y–but–not–Z confidence rating for this case is $(1-\phi(Z))$. All three cases can be generalized to a common form, however. Let $\phi_1(Y)$ be $\phi(Y)$ if Y is a borderline case opposite Z, 1 otherwise. And let $\phi_2(Y)$ be $\phi(Y)$ if Y is a borderline case next to Z, 1 otherwise. Then all the formulas we have exhibited fit the scheme $\phi_1(Y) * (\phi_2(Y) - \phi(Z))$.

With this insight, we can extend our comparison to the three-dimensional case. The case is shown in Figure 7 for a color description k.

To calculate this probability mass we generalize Y–but–not–Z calculation to a case analysis in three

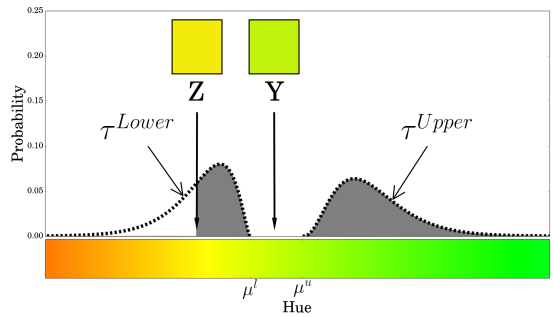


Figure 6: If Y is a clear case, we simply exclude Z, for probability $1-\phi(Z)$, here $1-0.38 = .62$.

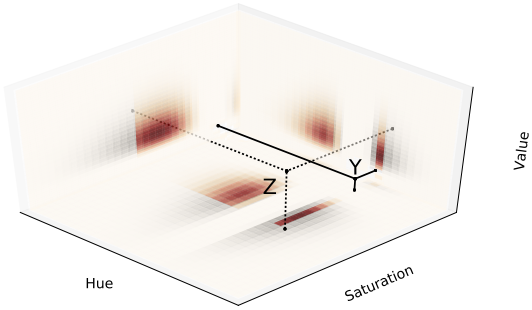


Figure 7: In the multidimensional case, solutions respect constraints from Y that are independent of Z, with probability $\phi_1(Y)$; they also select appropriate standards that affect both Y and Z, with probability $\phi_2(Y) - \phi_3(Z)$.

dimensions as shown in Equation 4.

$$\phi_1(Y) * (\phi_2(Y) - \phi_3(Z)) \tag{4}$$

In this equation, we generalize our notation to the general case as follows:

- $\phi_1(Y)$ is $\prod\phi(Y^d)$ over dimensions d where Y and Z are contrasting borderline cases
- $\phi_2(Y)$ is $\prod\phi(Y^d)$ over all other dimensions
- $\phi_3(Z)$ is $\prod\phi(Z^d)$ over all dimensions d

This expression is what we use in our implementation to calculate each color term's Y–but–not–Z confidence rating.

Given a confidence score, the evaluation is balanced by the availability of the color description.

Algorithm 1 The scoring function to compare two HSV tuples Y and Z for a single color term k

```
function SCORE(k,Y,Z)
    TermA ← 1
    TermB ← 1
    TermC ← φ_k^H(Z^H) × φ_k^S(Z^S) × φ_k^V(Z^V)
    for each dimension d in (H,S,V) do
        if Y^d is on opposite side from Z^d then
            TermA ← TermA × φ_k^d(Y^d)
        else
            TermB ← TermB × φ_k^d(Y^d)
        end if
    end for
    score ← TermA × (TermB − TermC)
    score ← score × α(k)
    return score
end function
```

For example, a common color term like *green* has a high availability, whereas a less frequent term, *British racing green*, has a much lower one. By weighting a term's score by its availability, we ensure that the program is less likely to generate rare color labels unless they clearly target a difficult distinction that the program needs to make.

With this score function complete, we arrive at the basic outline of our algorithm. The algorithm is shown in Algorithm 1. The *distinguish* function cycles through the dictionary, calculates the *Y*–but–not–*Z* confidence for each term *k*, and returns the results in sorted order. In the cases in which *k* describes *Z* better than it describes *Y*, the function will evaluate to a negative number. Such cases are rejected—given our model, the terms cannot describe *Y* without also describing *Z*.

5 Results

We have created an interactive visualization that allows viewers to confirm the qualitative properties of our model for themselves. Figure 8 shows a screenshot of the visualization.

Users click on either of the two color swatches on the left to select colors, which are passed to the program as two HSV triplets. The middle column then displays a list of color terms associated with those swatches; this is context-independent data pulled directly from LUX. Terms are displayed in two colors: terms that are generally good descriptions of the target color but are bad at distinguishing it from its alternative are grayed out. For example, *light green* is grayed out at the top in Figure 8, because it's such a good description of the lower swatch. The column on the right then displays the results of the generation model for the two colors. Typically, no term appears in both lists—as is true in Figure 8—because it's rare to find cases like Figure 1 where there are two plausible, competing ways to refine the meaning of a color term so as to fit one color but not the other.[2] Results are ranked by normalized confidence values; colors move up in the rankings when they more precisely distinguish the target color from its alternative. For example, *pale green* and *yellow-green* overtake the more general *spring green* as descriptions of the lower color in Figure 8.

[2]Our model does recognize a surprising difference between *lime* and *lime green* in Figure 8. This isn't a fluke: the same difference shows up in CSS color definitions for example. We suspect that *lime green* evokes the peel of the fruit but *lime* is named for the juice.

Because the colors in Figure 8 are so close, context has a strong effect in selecting differentiating descriptions. As the two colors get further apart, there's less probability mass assigned to interpretations that categorize them the same way. Under these circumstances, the differentiating color terms converge to the color terms predicted by the generic model. This recalls the heuristic of Baumgaertner et al. (2012) that basic color terms are used unless needed to distinguish. In other words, our model produces marked descriptions only when coarser terms are less reliable in distinguishing the two colors, so they are necessary to achieve the communicative goal of distinguishing the two colors. This recalls the "small gaps" constraint of van Deemter (2006).

As a first step towards quantifying the performance of the model, we got the data collected by Baumgaertner et al. (2012). They showed subjects color swatches in arrays of four, and asked subjects either to identify a particular target swatch in words (as director) or to pick the swatch that best fit a verbal description (as matcher). At issue was the ability of human matchers or various algorithms to find the original target swatch (the correct swatch) given directors' descriptions. People's success in these tasks depends on how difficult it is to distinguish the alternatives. Because problems are so variable and task dependent, there can be no universal benchmark of performance in identifying colors, but the results are helpful in understanding what we have accomplished and where further research is necessary.

Baumgaertner et al. (2012) report an analysis of 29 judgments about the interpretation of color descriptions in context across a range of difficulty levels. Their baseline algorithm, which interprets colors based on the nearest focal value in RGB space, links 23 of them to the swatch the director was instructed to describe. Of the remainder, three represent clear problems with their system. Our system, by contrast, gets all these 26 correct. The remaining three cases raise the same problems for both approaches. There seems to be one case of human error: the director is signaled to describe a brown swatch but produces *blueberry*, apparently describing the adjacent purplish-blue swatch. And two are cases of sparse data: the items *deep grey blue* and *dull salmon pink* fall out of the frequent vocabulary of Munroe's data set. The two out-of-vocabulary cases arise in the most difficult setting,

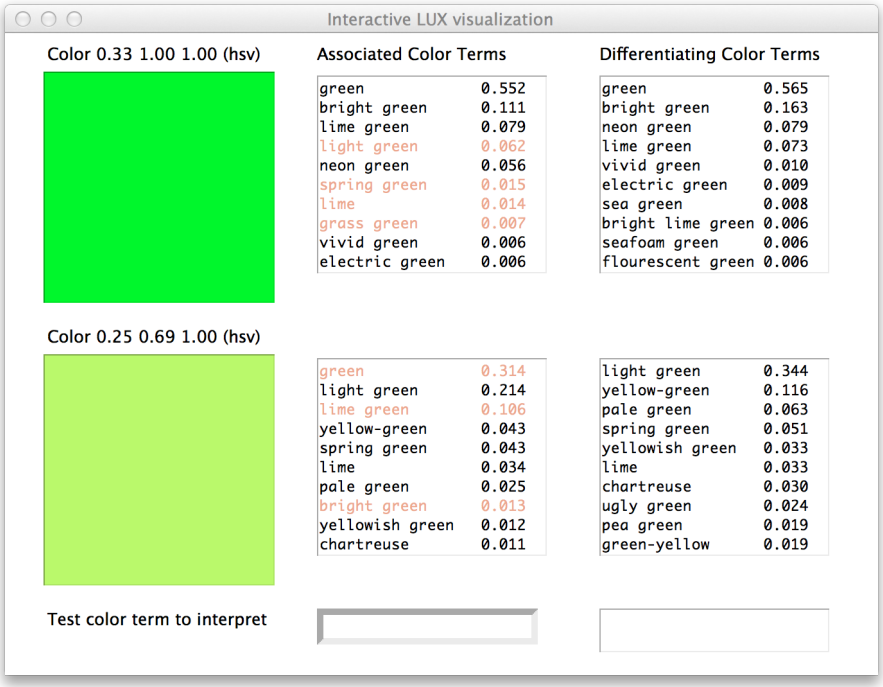


Figure 8: A screenshot of our interactive visualization, contrasting two shades of green. The system's descriptions emphasize the greater saturation and greener hue of the top color, and the lower saturation and yellower hue of the bottom color.

where directors must use low frequency terms to describe closely related colors; we get 71% right while human matchers recover the swatch signaled to the human director only 78% of the time.[3] Thus, we conclude that we need larger and more targeted data sets to distinguish the performance of our new algorithm from that of people.

Baumgaertner et al. (2012) 29 key examples are drawn from a larger elicitation experiment that produced 196 different tokens, again across a range of conditions. Our system resolves 152 correctly as written. Another 28 are out of vocabulary but closely related to terms the system would resolve correctly (differing in spelling, comparative or superlative morphology, hedges, paraphrases or other lightweight modifiers). The system gets 8 wrong as written (again, this seems to include several cases of human error); 6 are out of vocabulary and closely related to terms that the system would get wrong; and 2 are completely different from any of our vocabulary items. All the system errors are on low frequency items in situations with close distractor colors, where we've seen people also have difficulty. We were unable to find patterns of systematic error in our system.

[3]Interestingly, our system correctly resolves the alternative items *dark grey blue* and *salmon pink* in these cases. If we can deal with the productivity of low frequency descriptions, we see no obstacle to matching or even exceeding human performance.

6 Conclusion

We have explored a problem solving approach to the use of vague language. We have presented the theoretical rationale for our approach, described a broad-scale implementation, and offered a preliminary empirical evaluation.

Our work is pervasively informed by previous work on the semantics and pragmatics of dialogue. But we have not deployed or evaluated our work with interactive language use. That's an obvious and important next step.

We're excited by opportunities our work brings to assess the role of linguistic knowledge and rational problem solving in conversation. If successful, these efforts will lead to better interactive systems. But even if not, we think they will help to characterize speakers' interactive strategies, and thus to pinpoint the distinctive mechanisms that support meaning making in dialogue.

Acknowledgments

This work was supported in part by NSF DGE-0549115. Thanks to the SEMDIAL reviewers for helpful comments.

References

[Anderson1991] John R. Anderson. 1991. The adaptive nature of human categorization. *Psychological Review*, 98(3):409.

[Barker2002] Chris Barker. 2002. The dynamics of vagueness. *Linguistics and Philosophy*, 25(1):1–36.

[Baumgaertner et al.2012] Bert Baumgaertner, Raquel Fernández, and Matthew Stone. 2012. Towards a flexible semantics: colour terms in collaborative reference tasks. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 80–84. Association for Computational Linguistics.

[Clark and Wilkes-Gibbs1986] H. H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

[Gärdenfors2000] Peter Gärdenfors. 2000. *Conceptual Spaces*. MIT.

[Jäger2010] Gerhard Jäger. 2010. Natural color categories are convex sets. In Maria Aloni, Harald Bastiaanse, Tikitu de Jager, and Katrin Schulz, editors, *Logic, Language and Meaning - 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, volume 6042 of *Lecture Notes in Computer Science*, pages 11–20. Springer.

[Kennedy and McNally2010] Chris Kennedy and Louise McNally. 2010. Color, context and compositionality. *Synthese*, 174(1):79–98.

[Kennedy2007] Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45.

[Kyburg and Morreau2000] Alice Kyburg and Michael Morreau. 2000. Fitting words: Vague words in context. *Linguistics and Philosophy*, 23(6):577–597.

[Larsson2013] Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

[Lassiter2009] Daniel Lassiter. 2009. Vagueness as probabilistic linguistic knowledge. In Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz, editors, *Vagueness in Communication - International Workshop, ViC 2009, held as part of ESSLLI 2009, Bordeaux, France, July 20-24, 2009. Revised Selected Papers*, volume 6517 of *Lecture Notes in Computer Science*, pages 127–150. Springer.

[Lewis1979] David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(3):339–359.

[Ludlow2014] Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford University Press, Oxford.

[McMahan and Stone2014] Brian McMahan and Matthew Stone. 2014. A Bayesian approach to grounded color semantics. Manuscript, Rutgers University.

[Roy and Reiter2005] Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artif. Intell.*, 167(1-2):1–12.

[Steels and Belpaeme2005] Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language. A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–529.

[Stone and Webber1998] Matthew Stone and Bonnie Webber. 1998. Textual economy through close coupling of syntax and semantics. In *Proceedings of International Natural Language Generation Workshop*, pages 178–187.

[Thomason1990] Richmond H. Thomason. 1990. Accommodation, meaning and implicature. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 325–363. MIT Press, Cambridge, MA.

[van Deemter2006] Kees van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

[Williamson1996] Timothy Williamson. 1996. *Vagueness*. Routledge, London.

[Young et al.2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Learning to understand questions

Sara Moradlou
LLF (UMR 7110) & LabEx-EFL
Université Paris-Diderot (Paris 7)
Sorbonne Paris Cité, Paris, France
sara.moradlou@gmail.com

Jonathan Ginzburg
CLILLAC-ARP & LabEx-EFL
Université Paris-Diderot (Paris 7)
Sorbonne Paris Cité, Paris, France
yonatan.ginzburg@univ-paris-diderot.fr

Abstract

Our aim in this paper is to characterise the learning process by means of which children get to understand questions. In contrast to the acquisition of *production* of questions, an area which has a long history, the emergence of question comprehension is largely uncharted territory. we limit our attention in this paper to *wh*–interrogatives, since generally there is overt evidence for their understanding before other types of questions such as polar questions. The general idea we follow is that the child learns to understand questions interactively, as there is a long period of "training" during which the carer asks questions and answers them himself. Since the answers can be understood by the child, given sufficient exposure the child deduces an association between the pre-answer utterance and a question. Nonetheless, the process as we describe it here assumes a number of very strong priors. In particular, we will be assuming for some stages of the process that the child is attuned to a very simple *erotetic logic*—a logic which given certain assumptions allows one to deduce questions. We provide evidence for our model based on classifying interactions between a child and her parents in the multimodal Providence corpus from CHILDES.

1 Introduction

Our aim in this paper is to characterise the learning process by means of which children get to understand questions. In contrast to the acquisition of *production* of questions, an area which, as we discuss in section 2, has a long history, the emergence of question comprehension is largely uncharted territory, to the best of our knowledge.

We equate the comprehension of a question with the ability to provide an answer that concerns the question (in the sense of aboutness answerhood (Ginzburg, 2010), hence no requirement that such an answer be true.).

The general idea we follow is that the child learns to understand questions interactively, as there is a long period of "training" during which the carer asks questions and, receiving no answer, answers them himself. Since the answers can be understood by the child, given sufficient exposure the child deduces an association between the pre-answer utterance and a question. Nonetheless, the process as we describe it here assumes a number of very strong priors. In particular, we will be assuming for some stages of the process that the child is attuned to a very simple *erotetic logic*—a logic which given certain assumptions allows one to deduce questions (Wiśniewski, 2013a). This means that one needs to distinguish between the task of question acquisition and the more purely cognitive task of the emergence of erotetic reasoning; of course a similar delimitation is required to distinguish the emergence of beliefs and the understanding of the contents of declarative utterances.

In terms of data, we limit our attention in this paper to *wh*–interrogatives, since generally there is overt evidence for their understanding before other types of questions such as polar questions—a potentially interesting puzzle for most theories of questions where the latter are somehow simpler entities. However, we do discuss which of the learning strategies we consider scales up to polar questions, and will extend the empirical coverage to polars in an extended version of this paper.

Beyond the intrinsic interest of the topic of the acquisition of questions, we think that this is a topic that can ultimately offer grounds for selecting among existing theories of questions on the grounds of learnability.

The structure of the paper is as follows: in sec-

tion 2 we survey previous work on questions, on the acquisition of the production of questions, and on Bayesian learning. In section 3 we discuss the games by means of which we hypothesise questions get learnt. Section 4 provides the empirical evidence evaluating the plausibility of our approach.

2 Previous Work

2.1 Questions and Semantics

In considering how questions are acquired, we need to settle on a representation of the target entity, viz what a question is. Although there has been much work in formal semantics on the meaning of interrogatives (for surveys see e.g., (Groenendijk and Stokhof, 1997; Ginzburg, 2010; Wiśniewski, 2013b)), as Wiśniewski says 'No commonly accepted theory of questions has been elaborated so far.' The questions literature has not addressed the issue of how questions might be acquired, nor the cognitive plausibility of the semantic entity a given theory assumes as an interrogative denotation. On grounds of cognitive tractability, from among currently influential views, neither the partition theory, where a question is seen to be a partition of the set of possible worlds (for detailed motivation see (Groenendijk and Stokhof, 1997)), nor the inquisitive semantics view, where a question is seen to be a set of sets of worlds (see (Wiśniewski, 2013b) for detailed discussion) can be candidates (though one cannot rule out the possibility of cognitively tractable versions being formulated.). We will assume a view of questions as propositional functions, a view apparently initiated by (Ajdukiewicz, 1926), developed significantly in (Kubinski, 1960), and subsequently shared and further developed by a number of different approaches (Krifka, 2001). We adopt an implementation of this view within the framework of Type Theory with Records (Cooper, 2010). Specifically, it will be convenient to think of questions as records comprising two fields, a situation and a function (Ginzburg et al., 2014). The role of *wh*-words on this view is to specify the domains of these functions; in the case of polar questions there is no restriction, hence the function component of such a question is a constant function. (1) exemplifies this for a unary 'who' question and a polar question:

(1) a. $Who = \begin{bmatrix} x_1 : Ind \\ c1 : person(x_1) \end{bmatrix}$; $Whether = Rec$;

b. 'Who runs' $\mapsto$ $\begin{bmatrix} sit = r_1 \\ abstr = \lambda r{:}Who(\begin{bmatrix} c : run(r.x_1) \end{bmatrix}) \end{bmatrix}$;

c. 'Whether Bo runs' $\mapsto$ $\begin{bmatrix} sit = r_1 \\ abstr = \lambda r{:}Whether(\begin{bmatrix} c : run(b) \end{bmatrix}) \end{bmatrix}$

Given this, the following relation between a situation and a function is the basis for defining key coherence answerhood notions such as resolvedness and aboutness (weak partial answerhood (Ginzburg, 2010)) and question dependence (cf. erotetic implication,(Wiśniewski, 2013b)):

(2) s *resolves* q, where q is $\lambda r : (T_1)T_2$, (in symbols $s?q$) iff **either**
(i) for some $a : T_1$ $s : q(a)$,
or
(ii) $a : T_1$ implies $s : \neg q(a)$

2.2 The emergence of wh-interrogative production

There appears to be a relatively robust order of acquisition of the production of wh-words in questions reported for a variety of languages, in which 'what' and 'where' (and their cross-linguistic equivalents) are acquired before other wh-words (e.g., 'why', 'how' and 'when') (Brown and Hanlon, 1970; Bloom et al., 1982). Bloom and collaborators proposed a complexity-based account. On this line, the first wh-questions to emerge are wh-identity questions—questions that ask for the identities of things or places. These are suggested to occur with what Bloom et al. term the 'relatively simple' 'what' and 'where', and should occur primarily with the copula. Later on, the *wh*-words, which now also include 'who', are envisaged to start occurring with a greater variety of main verbs (e.g. 'Where has he gone?', 'What are you doing?'). There have also been more recent alternative accounts of such phenomena in terms of input frequency (see (Theakston et al., 2001; Rowland et al., 2003), and references therein).

2.3 Bayesian learning and semantics

Recent years has seen the emergence of formal accounts of deep semantic learning rooted in a Bayesian approach to cognition.

(Piantadosi et al., 2012a; Piantadosi et al., 2012b) propose an approach which they apply to the learning of, respectively, numeral systems and quantifier expressions. The general strategy is to use the λ-calculus as a means for developing a hypothesis space (a *language of thought* for the learner, in the authors' words.). Restricting ourselves to the numeral case: a space of functions from sets to number words is defined (including a function representing knowledge that singleton sets can be counted by the word 'one', doubleton sets by 'two' and fails on any other type of set, a function that partitions all sets into either 'one' or 'many' etc). The crucial ingredient concerns how the learner chooses among these hypotheses: a probabilistic model is constructed built on the idea that the learner should attempt to trade-off two desiderata. On the one hand, the learner should prefer a lexicon having a short description in the language of thought. On the other hand, the learner should find a lexicon which can explain the patterns of usage seen in the world. Balancing these requirements is effected by using Bayes' rule.

Frank et al. (2009) attempt to synthesise two approaches to word learning, one based on recognition of speaker intention and one based on cross-situational learning. The model constructed consists of a set of variables representing the word-learning task and a set of probabilistic dependencies linking variables representing the lexicon of the language being learned, the referential intentions of the speaker, the words uttered by the speaker, and the learner's physical context at the time of the utterance. The physical context of an utterance is identified as the set of objects present during the utterance, the speaker's referential intention as the object or objects he or she intends to refer to, and the lexicon as a set of mappings between words and objects. Using an observed corpus of situations—utterances and their physical context—the model works backward using Bayesian inference to find the most likely lexicon.

We hypothesise these methods could be extended to learning the meanings of wh words. However, in both cases what we have is batch learning of sets of lexical items, which as the authors acknowledge makes no reference to the interaction between parent and child, so falls short of a theory of the process in which acquisition emerges.

3 Modelling

The narrative We consider three potential games of increasing complexity for learning questions. The first one will lead to success but can only enable the learning of a small class of questions. The second game is significantly more general, but still quite restricted. The third one is yet more general (though not fully sufficient for learning questions), but here success is far less clear. We hypothesise that this sequence can be used to explicate the order of comprehension of questions. To what extent this hypothesis is vindicated is discussed in section 4.

3.1 Salient Object Identification (SOI)

Priors understand 'that', shared gaze/deixis, predication

The game: training phase while sharing gaze at an object the parent asks a question that involves the child identifying the object or the object's location. The parent offers the child the opportunity to answer and when no response is forthcoming, the parent offers a name, attribute, or deictic gesture.[1]

Examples[2]

(3) a. [Mother turns page to reveal page with mirror on it.]: who's that? who's that ? huh ? can you see ? rabbit.

b. [Mother walks Big Bird up] who's that? who's that? is that Big_Bird ?

Rationale In the training phase the child is unsure how to respond: as far as a language like English that has *wh*–fronting, the initial hypothesis (given her existing lexicon of NP meanings) is that 'what' or 'who' is referential; this conflicts with the normal structure of copular sentences (*Bo is that, *The ball is this). Still, in the absence of an alternative, some initial high probability has to be assigned to the hypothesis that these words are referential. Since the range of questions asked is small, it is feasible to be making and retaining hypotheses about the meaning of this (type of) unclear utterance. Once the parent provides the relevant answer, the child understands the answer

[1] We are assuming that turn taking is being acquired independently, as a tool used in a variety of move types, indeed not just for linguistic purposes.

[2] All the examples in this section are taken from the Rollins corpus, (Rollins, 2003).

since the word is chosen to be known to the child and it predicates of the entity in visual focus.

It is not clear though that there is anything in this interaction to argue *against* the hypothesis that the 'wh' words refer to the entities picked out,[3] Nonetheless, given sufficient exposure to this game, the child gets habituated to associating with the utterance of the interrogative utterance the predication of a property of the salient entity in the situation and this process does involve the child considering various possible properties for classifying that entity. In other words, a data structure individuated by a situation and a function, as in (1b). So there is a holistic content associated with the interrogative utterance, not one built up compositionally.

weaknesses This game underdetermines answerhood since neither negative nor quantified answers will be encountered. Furthermore, it will not scale up to learning other types of questions, most obviously polar questions.[4]

3.2 Erotetically plausible questioning

Priors understand 'that', shared gaze/deixis, predication, an erotetic inference capability (Wiśniewski, 2013a)—awareness that certain situations raise questions: when shown an object, the question will be: who/what is that?; when an object disappears, the question will be: where is SO?; seeing animal, what noise does it make? seeing an object: what things can it do? etc. We call questions deduced in this way in context *erotetically plausible questions* (EPQ). The erotetic capability assumed is a parameter of the game—different games will ensue with the assumption of different erotetic capabilities.[5]

[3]There is Eve Clark's contrast principle (Clark, 2002) which is potentially of some help, given the need to distinguish 'what' or 'who' from 'that'. But given they do differ from 'that' via their associated restrictions, it is not obvious that would be sufficient.

[4]There are clearly polar question oriented games, such as those where a child gets to respond by shaking their head as a negative response. What is important to ascertain is how general the notion of negation used there is, to what extent this is distinct from expressing a negative volition. We hope to investigate this point in subsequent work.

[5]An anonymous reviewer for SemDial cautions us from identifying too closely the notion of erotetic inference capability with that associated with e.g., Wiśniewski's IEL. This is of course a reasonable point, though in pointing towards formalisms like IEL our intention is to highlight the apparent use of reasoning that employs questions, not solely propositions. IEL is in any case a rather general framework, consistent with many distinct conceptions of semantics and reasoning.

The game: training phase in a situation s the parent asks a question that is EPQ in s. The parent offers the child the opportunity to answer and when no response is forthcoming, the parent offers an answer.

Examples

(4) a. [Mother pulling hair from rattle]: where is all this hair coming from?

b. [Mother removes big bird] Where did Big Bird go? [pulls big bird up into line of sight] peek a boo.

Rationale The EPQ game generalises SOI by allowing a wider range of questions, emphasising the likelihood of the question in context; it can, in principle, scale up to polar questions (e.g., pressing a balloon from both sides raises the issue of whether it will burst.) and a wider range of answers. Understanding the answer is less deterministic than with SOI since a given context could be compatible with a number of questions arising. But, once again, a small number of possible questions and sufficient training potentially habituate the child to associate situations which trigger erotetic inferences with questions in a holistic way.

weaknesses There is the potential for mismatch between the child's internal erotetic capabilities and those associated with the natural language used. The range of potential questions that can be learnt in this way is still severely restricted.

3.3 Situational Description Games

Priors Similar to EPQ games.

The game: training phase In a situation s the parent asks questions about properties of objects in the observed situation, described using words the child knows. The parent offers the child the opportunity to answer and when no response is forthcoming, the parent offers an answer.

Examples

(5) a. [Mother looks at book]: what kind of colors do we have here ? [puts book on tray] look there's purple. that's Mot [= mommy's] favorite color. and pink. and blue.

b. [Child holding car] what's on this car ? [grabs other side of car Chi has in hand and turns it over .] this car has a butterfly sticker on it.

Rationale This game can be extended to cover an unrestricted range of questions (though of course by no means the full range of NL questions.).

weaknesses There is no guarantee that the child will understand the answer, hence there is no guarantee that learning of a given interrogative meaning will succeed. But assuming the child has been well trained with EPQ, the child will habituate to associate interrogatives with a wider range of questions than EPQ.

3.4 Formal characterisation of the games

Each of these games can be characterised formally as a *genre* in the sense of (Larsson, 2002; Ginzburg, 2010)—an interactional sequence with restricted subject matter. We demonstrate how to do so in the extended version of this paper.

4 Data

We randomly sampled and selected 20 wh-questions of each file (31–48% of all wh-questions present in the files[6]) from early files of Naima of Providence corpus (Demuth et al., 2006). These questions were annotated for their form, child's response, mother's follow-up, evaluation of child's answer, and the semantic model that describes them best (SOI, EPQ, SDG, as discussed previously).

4.1 Caregiver's questions

Naima's parents asked 'what' and 'where' questions most frequently (see Table 1). As shown in Table 2, the SOI question interactions almost solely occur with copular structures, whereas the other more complex games appear with a wider range of constructions. We did not find any evidence that caregivers present children with the games we discuss above sequentially (i.e. frequency of the games did not change in favor of more complex ones over time.). One could argue however, that the relatively simple, almost fixed, structure[7] of questions in SOI makes those questions more tractable and bootstraps the learning process.

[6]Wh-questions comprised 24.4–30.3% of all questions (including polar questions, choice questions, etc.).

[7]We take the word type following the wh-word to be a reasonable proxy for measuring structural complexity.

	which	who	who else	where	what	what else
SOI	1	8	0	4	11	0
EPQ	0	1	0	3	10	1
SDQ	0	1	1	23	23	4
OTH	0	0	0	1	8	0
total	1	10	1	**31**	**52**	5

From files 1, 3, 5, 7, and 9 of Naima

Table 1: Frequency of wh-words with the semantic class of the question

	SOI	EPQ	SDG	other
—	0.04	0.07	0.08	0.11
AUX	0.04	0.07	0.21	0
MOD	0	0.07	0	0.11
COP	**0.92**	0.40	0.56	0.33
DO	0	0.27	0.13	0.44
V	0	0.13	0.02	0

From files 1, 3, 5, 7, and 9 of Naima

Table 2: Percentages of forms following wh-word in parental questions and their semantic class

4.2 Children's answers

The annotator judged the correctness of child's response with respect to the question and the situation and tagged the instances as Correct (C), Type Correct (TC), Incorrect (IC), and Not Attempted (NoA).

We argued above that SOI and EPQ questions are easier for child to answer compared to SDG. Table 3 shows that SOI and EPQ questions get answered more often and might therefore be easier to learn.

Naima was more likely to attempt answering (irrespective of the correctness of the answer) SOI and EPQ questions compared to SDG and these attempts also increased by age ($Pr(> |t|)s < .05$). We observed the same patterns for the correctness of the answers (i.e. SOI and EPQ questions were answered more correctly (on the scale of NoA <

Sem	answered C/TC (%)	total (#)
SOI	58	24
EPQ	60	15
SDG	38	52
Other	12	8

From files 1, 3, 5, 7, and 9 of Naima

Table 3: Percentage of questions answered by child

	Sem (% answered C/TC)				
Age	SOI	EPQ	SDG	Other	total
11.28	60	33	14	0	30
12.28	67	67	45	–	55
13.25	33	–	33	0	30
14.23	100	100	40	0	45
15.12	67	100	57	33	60

From files 1, 3, 5, 7, and 9 of Naima
Ages in month.days

Table 4: Percent questions answered by child over age

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.752	0.580	-3.020	0.003 **
SemEPQ	1.409	0.667	2.113	0.037 *
SemSOI	1.595	0.584	2.731	0.007 **
SemOTH	-2.040	1.123	-1.817	0.072 .
Age	0.256	0.091	2.803	0.006 **

Signif. codes: 0.0001 *** 0.001 ** 0.01 * 0.05 . 0.1
Formula : CEval ∼ Sem + Age Intercept terms (reference levels): No Answer, SDG, and Youngest age.

Table 5: Best fitting model of evaluation of child's answer

IC < TC < C) than SDG questions and age of the child showed a positive main effect on this correctness. See Table 5).

We also annotated the type of answer Naima provided to her mother's question as "ShortAns" when she responded with a single word utterance that was relevant to the question, and as "ActAsAns" when she responded to the question with a relevant action. We coded utterances that did not pertain to the question with "IrRel" and no attempt to answer as "NoA". Our Fisher's exact test revealed that a child's answer to a question significantly differed by its semantic class (p-value = 0.041). Using correspondence analysis, Figure 1 illustrates the trends in this correlation: Child's ShortAns cooccurs with SOI, and to a lesser degree with EPQ. This was expected from analyses of answer attempts and answer correctness discussed earlier.

4.3 Mother's follow-up

Table 6 summarises our annotation schema for mother's follow-up utterances along with percentages of their occurrence in the data sampled from Providence (Demuth et al., 2006). The child's answers to the questions correlated significantly with

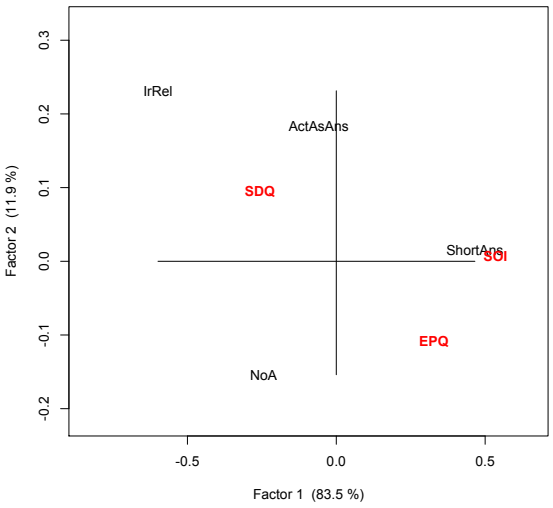


Figure 1: Correspondence analysis graph for child's answer and semantic class of the question

the mother's follow-up utterance (p-value = 5.24e-06). As indicated in Figure 2, the mother's IrRel follow-up is positively correlated with the child's IrRel; this is most likely due to a shift of topic or attention in the conversation. When the child gives no answer (NoA), the mother proceeds with reformulations (Rfl) or repetitions (Rpt) of the same question, or asks a new related question (RNQ). The child's answers (ShortAns and ActAsAns) on the other hand get meaningful feedback from the mother (ShortAns, SentAns:Simple, YES).

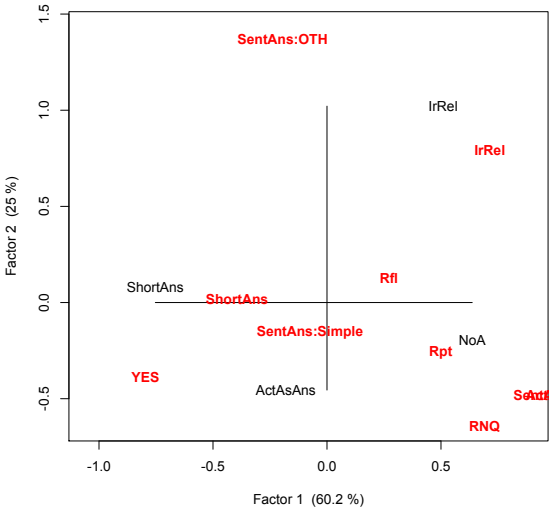


Figure 2: Correspondence analysis graph for child's answer and mother's follow-up

Mother's follow-up	%		Example
ShortAns	13	Short Answer	MOT: what is it? MOT: lego.
SentAns:Simple	18	Simple Sentential Answer	MOT: what's that? CHI: yyy dog. MOT: that's a little dog.
SentAns:PQA	4	Polar Question Answer	MOT: who's that? MOT: is that the doctor?
SentAns:OTH	2	Other sentential answer	MOT: what's that? CHI: shirt[?] shirt[?] MOT: it looks like pants to me but that's close.
ActAsAns	1	Action as answer	
Rfl	15	Reformulation	MOT: what's that? CHI: yyy. MOT: you know what that is?
Rpt	14	Repetition	MOT: where's dolly Naima? MOT: where's dolly?
RNQ	4	Related New Question	MOT: where's pipo? MOT: what's he doing?
IrRel	10	Irrelevant utterance	
RCA	10	Repeat Child's Answer	MOT: where'd [: where did] it go? CHI: down. MOT: down.
YES	9	Acknowledge Child's Answer	MOT: who else do we see in that picture? CHI: pony. MOT: yeah.

Percentages and examples from Naima, files 1, 3, 5, 7, and 9.

Table 6: Mother's follow-up utterances

4.4 Earlier input

We also looked at 18 files from the Rollins corpus (Rollins, 2003) to investigate to what extent caregivers provided answers to their own questions during the stage where children didn't produce any answers at all.[8] Table 7 indicates that even in the earlier stages caregivers answer about half of their own questions.

We did not find any significant effect of age on question words or mother's follow-up. Individual differences however, were significant for question word (X-squared $= 333.39$, $df = 63$, p-value $< 2.2e-16$). The numbers of 'what' and 'where' questions were significantly different for different mothers ($Pr(>|z|) < 0.01$)[9].

The complexity of the question forms, as measured by the second word[10], changed significantly with children's age with individual differences accounted for as random effects[11].

5 Conclusions and Future Work

In this paper we have offered a sketch of a theory of the emergence of question comprehension by children, within a type theoretic view of questions as situationally relativized propositional functions. We have outlined how this might happen with reference to certain restricted interaction sequences between parent and child, tying this to ease of classification of situations and erotetic inference capability that children develop. The data we present from the interactions of one child in the Providence corpus with her parents offers encouraging indications that the notions of question complexity we postulate are on the right track.

An important component that remains to be spelled out is the probabilistic reasoning underlying the various habituation states we have conjectured.

[8] Out of 422 questions, only 7 were answered to by children; only 2 of those answers were verbal.

[9] Generalised linear model with mother as dependent variable and question word as predictor.

[10] Words occurring right after question word were of the types: AUX, COP, MOD, DO, and V.

[11] Generalised linear mixed model Formula: $age \sim SecondWord + (1|name)$

	Children (% Mother's follow-up)									
Mother's follow-up	cb	di	ds	hc	im	jw	me	nb	sx	mean
ActAsAns	5.6	5.9	2.5	14.0	1.6	6.9	2.0	14.0	8.1	6.73
IrRel	11.0	5.9	25.0	14.0	11.0	17.0	13.0	28.0	16.0	15.7
Rfl	11.0	0.0	2.5	14.0	20.0	24.0	25.0	10.0	19.0	13.9
Rpt	11.0	24.0	28.0	32.0	16.0	14.0	20.0	12.0	14.0	19.0
SentAns:OTH	5.6	8.8	2.5	4.5	3.3	10.0	5.0	3.8	8.1	5.73
SentAns:PQA	11.0	32.0	0.0	9.1	25.0	6.9	15.0	5.0	11.0	12.8
SentAns:Simple	22.0	12.0	30.0	0.0	11.0	10.0	12.0	7.5	11.0	12.8
ShortAns	22.0	12.0	10.0	14.0	11.0	10.0	8.9	20.0	14.0	13.5
Answered	66.2	70.7	45.0	41.6	51.9	43.8	42.9	50.3	52.2	51.62
Answered verbally	60.6	64.8	42.5	27.6	50.3	36.9	40.9	36.3	44.1	44.9

Ages 9 and 12 months of nine children from Rollins corpus.

Table 7: Distribution of Mother's follow-up to her own questions

Acknowledgments

We would like to thank Eve Clark for much useful discussion, Nicholas Asher for a question which set us away from a previous erroneous path, three anonymous reviewers for SemDial 2014 and Eve Clark for very helpful comments on a previous version of this paper. This work has been funded by a doctoral fellowship for SM from the LabEx–EFL (ANR/CGI).

References

K. Ajdukiewicz. 1926. Analiza semantyczna zdania pytajnego. *Ruch Filozoficzny*, 10:194b–195b.

Lois Bloom, S. Merkin, and J. Wooten. 1982. Wh-questions: Linguistic factors that contribute to the sequence of acquisition. *Child Development*, 53:1084–1092.

R. Brown and C. Hanlon. 1970. Derivational complexity and order of acquisition. In J.R. Hayes, editor, *Cognition and the Development of Language*, pages 11–53. John Wiley.

Eve Clark. 2002. Making use of pragmatic inferences in the acquisition of meaning. In D. Beaver, S. Kaufmann, B. Clark, and Luis Casillas, editors, *The construction of meaning*, pages 45–58. CSLI Publications, Stanford, CA.

Robin Cooper. 2010. 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*. Elsevier.

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

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585.

Jonathan Ginzburg, Robin Cooper, and Tim Fernando. 2014. Questions and propositions in type theory. In Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin, editors, *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics*, pages xx–yy, Gothenburg, April.

Jonathan Ginzburg. 2010. Questions: Logic and interaction. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Linguistics, 2nd Edition*. North Holland, Amsterdam.

Jeroen Groenendijk and Martin Stokhof. 1997. Questions. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Linguistics*. North Holland, Amsterdam.

M. Krifka. 2001. For a structured meaning account of questions and answers. *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, 52:287–319.

T. Kubinski. 1960. An essay in the logic of questions. In *Proceedings of the XIIth International Congress of Philosophy (Venetia 1958)*, volume 5, pages 315–322, Firenze. La Nuova Italia Editrice.

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

ST Piantadosi, JB Tenenbaum, and ND Goodman. 2012a. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2):199–217.

ST Piantadosi, JB Tenenbaum, and ND Goodman. 2012b. Modeling the acquisition of quantifier semantics: a case study in function word learnability. *mit.edu*.

Pamela Rosenthal Rollins. 2003. Caregivers' contingent comments to 9-month-old infants: Relationships with later language. *Applied Psycholinguistics*, 24:221–234, 6.

Caroline Rowland, Julian Pine, Elena Lieven, and Anna Theakston. 2003. Determinants of acquisition order in wh-questions : re-evaluating the role of caregiver speech. *Journal of Child Language*, 30:609–635.

Anna Theakston, Elena Lieven, Julian Pine, and Caroline Rowland. 2001. The role of performance limitations in the acquisition of verb argument structure. *Journal of Child Language*, 28:127–152.

Andrzej Wiśniewski. 2013a. *Questions, Inferences, and Scenarios*. College Publications, London, England.

Andrzej Wiśniewski. 2013b. The semantics of questions. In Chris Fox and Shalom Lappin, editors, *Handbook of Contemporary Semantic Theory, second edition*, Oxford. Blackwell.

Dynamic Intention Structures for Dialogue Processing

Charles L. Ortiz, Jr.
Natural Language and AI Laboratory
Nuance Communications
Sunnyvale, CA U.S.A.
charles.ortiz@nuance.com

Jiaying Shen
Natural Language and AI Laboratory
Nuance Communications
Sunnyvale, CA U.S.A.
jiaying.shen@nuance.com

Abstract

We examine personal assistance dialogues and argue that some form of constraint relaxation is necessary during dialogue processing as often only a subset of the constraints present in the intentional structure reflecting earlier parts of a dialogue can be satisfied in the context of a new utterance. We combine a fine-grained formal representation of intention with a non-monotonic consistency-based intention revision process to support a model of structured and evolving propositional content that leads to a flexible discourse segmentation process. The approach provides a bridge between models of rational agency, plan-based models of dialogue and theories of dynamic discourse semantics.

1 The problem

Plan-based approaches to dialogue processing structure a dialogue hierarchically into discourse segment purposes (roughly, intentions) and their interrelation (Grosz and Sidner, 1986). Such structures are incrementally elaborated as a dialogue unfolds and are referred to as *intentional structures*. They are grounded in agent models of collaboration and formal models of intention. The interpretation of a new utterance is given relative to whether it signals the start of a new segment, contributes to an existing one or completes it. The notion of contribution to a segment is given in terms of whether the sub-task (or its constraints) reflected by the utterance can play a part in the success of the task corresponding to the embedding segment. So, for example, in the task dialogues between experts and apprentices discussed in Grosz and Sidner (Grosz and Sidner, 1986; Lochbaum, 1998) an expert can suggest an action to an apprentice who will either execute it, ask for clarification or report obstacles.

Dialogues between a VPA and a user, however, differ in an important way from expert-apprentice task-based dialogues: a user typically provides some initial constraints on a task that it seeks help on after which the VPA will attempt to formulate a plan to satisfy those constraints. Often, however, either only a subset of those constraints can be satisfied or the user might change his mind on the set of constraints during the dialogue. Consequently, it may not be possible to accommodate a new utterance during the discourse segmentation process if one does so in terms of whether the interpretation of the utterance is consistent with the constraints so far articulated in the current segment. Rather, the system must relax some constraints to properly situate the utterance within the existing discourse structure and then continuing with the dialogue or providing assistance.

Consider the following example of a possible dialogue between a person and a virtual personal assistant (VPA) of the future in the context of a request for help in organizing a meeting with some friends after a conference session.

1. [User:] I want to plan a get together after the last session.
2. [System:] At what time?
3. [User:] 7pm.
4. [System:] OK.
5. [User:] Book a table at an Italian restaurant near the hotel and let Brian know.
6. [System]: Zingari is available at 7pm.
7. [User:] That's good.

Consider the following possible alternative user continuations in the highlighted contexts.

8a. And I'd like to include some good wine. (*Zingari does not have a good wine list. An alternative, Barbacco, does but it is farther away.*)

8b. Reserve a table at Chevy's instead. (*Chevy's is a Mexican restaurant.*)

8c. I decided that I want Spanish food.

8d. Actually, let's just go to a place for drinks. (*Zingari is not available for just drinks.*)

Consider each of these in the context of a SharedPlans (Grosz and Kraus, 1996) plan augmentation algorithm (Lochbaum, 1998). Lochbaum's algorithm determines the contribution of the interpretation of the current utterance, u, through a construction process which builds a complex recipe structure (action decomposition hierarchy) from simpler two-level recipes. In the last step, the algorithm checks for consistency: if the new constraints from u are satisfiable with those so far articulated in the recipe structure then they are combined, otherwise the algorithm fails and the user is alerted or queried.

Returning to the above examples, in each of the continuations the user introduces new constraints into the planning process[1] As it happens, each new constraint is inconsistent with the constraints communicated so far in the dialogue. In (8a), the system must engage in some constraint relaxation as otherwise the segment initiated by (5) would fail (and, hence, (8a) would have to be interpreted as part of the higher level segment: roughly, "I want to plan a get together with Brian and include some good wine"). This is to be expected and constitutes the reason the user needs assistance in the first place: he has no idea whether Zingari has a good wine list. The system may then try to find something a little farther away that meets all of the other constraints. In continuation (8b), one cannot simply delete the identity of the restaurant of Zingari and substitute that of Chevy's: there is also an inconsistency with a side-effect of the choice of Chevy's: the fact that it is a *Mexican* restaurant. In case (8c), the constraint that the restaurant be Italian is retracted which entails that the Zingari reservation be withdrawn. In case (8d), the method of setting up a get together is changed: the user decides to have drinks (nominally, at a bar) instead of going to a restaurant. However, a search should then not be constrained to an *Italian* bar. In all of these cases the binding for the objects in the second part of the action ("let Brian know") must depend on the different choices from the individual cases (so that if Zingari is picked, the VPA informs Brian and if the choice is changed to Barbacco the VPA informs Brian of the new location).

There are other concerns that these examples raise. Intention revision must be able to modify propositional content in a fine-grained way, rather than just deleting an inconsistent intention. For example, if we consider the first conjunct of (5), we would have something like:

$$intends(System, \exists x \exists t.occurs(book(x), t) \\ \wedge table(x) \wedge restaurant(x) \wedge italian(x))$$

Notice the embedded existential quantifier: the system has not fixed the identity of the restaurant or the time of the booking action. By utterance (8), however, those decisions have been made and the content within the scope of the above modal intention operator must be accessed and updated, without having to re-write the entire formula or delete the entire formula if the revising component is inconsistent with the intention: one would like to minimally modify the *contents* of the intention, unlike that in the belief revision literature. Dynamic Intention Structures (DIS), developed for modeling rational agents (Ortiz and Hunsberger, 2013; Hunsberger and Ortiz, 2008), addressed these problems. Whereas DRT makes use of a dynamic logic to deal with dynamic scoping of quantifiers, the theory of DIS's extends that idea to modalities with hierarchically structured content: the structure informs the consistency based revision procedure which lumps related elements, allowing incremental revision.

The similarity to DRT also addresses a perceived need, that has has been pointed out by others (Asher and Lascarides, 2003), to create a bridge between plan-based approaches and discourse *semantics* in a manner similar to approaches grounded in DRT.

2 Dynamic Intention Structures

We will present different forms of DISs that each serve different purposes. A *canonical* DIS is of the form $\langle V_c, T_c, int[\langle V_t, T_t, \langle\!\langle Id_r, A_r, V_r, T_r, E_r, C_r, S_r \rangle\!\rangle \rangle]\rangle$. V_c is a set of variables ("c" for "context") and T_c is a time point coinciding with the time of the intention. These external variables and time are existentially quantified in the translation to first order logic (FOL). V_t is a set of variables

[1] A reasonable segmentation would consist of a top-level segment for (1) and two sub-segments for (2)-(4) and (5)-(7). The DIS of Figure 1 lumps together the representations of (1)-(4). We do not delve into the precise mechanism behind the segmentation process as it does not bear directly on our presentation. A more detailed presentation would require a review of SharedPlans which are used to guide that process.

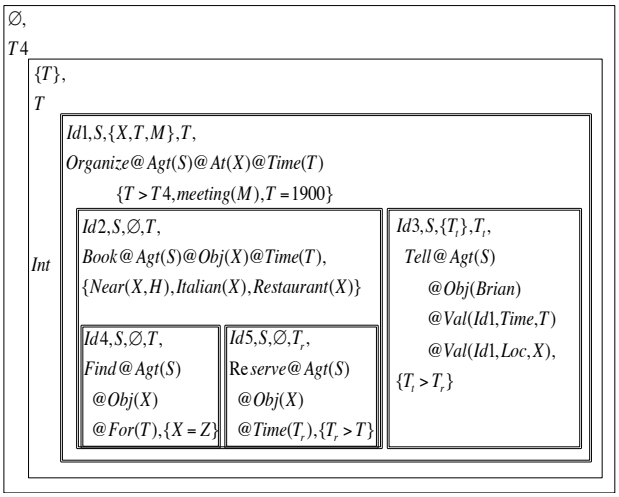


Figure 1: DIS after utterance 8.

and T_t is the time of the intended action We call each element within the double angle brackets ($\langle\langle \ldots \rangle\rangle$) a *plan tree node*. Each node has a unique identifier, Id_r ("r" for "root"); a possibly empty list of child nodes, S_r; an agent, A_r; an action, E_r; a set of local variables, V_r; and a set of constraints, C_r on actions and objects. Variables local to the intention are existentially quantified in the semantics, which allows one to express partially elaborated intentions such as "John intends to reserve *a room*" without necessarily fixing the identity of the room or the method for accomplishing the reservation.

The action representation makes use of an act-type constructor, @, which allows the construction of more complex act *types* from simpler ones (Ortiz, 1999). For example, suppose $drive@agt(A)@to(Boston)$ represents the act type of agent A driving to Boston. That act type could later be elaborated with further detail. For example, $drive@agt(A)@to(Boston)@on(Interstate95)$, might represent the act type of agent A driving to Boston via Interstate 95. In this way, the act-type constructor enables the representation of partially specified intentions without committing to a particular predicate arity—such as, $drive(Agent, Object)$.

To deal with partiality of action descriptions more systematically, the arguments to act-type modifiers will often be restricted to variables. For example, the preferred description of agent A driving to Boston would be $do(drive@agt(x)@to(y)) \wedge (x = A) \wedge (y = Boston)$. This technique has the advantage of enabling complex revisions to be performed simply by assigning or de-assigning values to variables.

DIS's, like DRSs, can be conveniently visualized using box notation. Figure 1 depicts the DIS obtained after utterances 1-8 (the constant h stands for the hotel). The representation is built up incrementally: the two outer boxes and the first box in the scope of the intention, labeled $Id1$, is a consequence of utterances (1-4). It says that at time $T4$,[2] the agent S (for system), intends to organize a meeting, M, at some yet to be specified location, X, (represented by the partially specified act-type term, $organize@Agt(S)@obj(M)@At(X)@Time(T)$, with constraint $meeting(M)$) at time $T = 1900$ (7 PM) in the future.

Utterances (5-7) lead to the full structure depicted in Figure 1: sub-actions corresponding to boxes $Id2$ and $Id3$ are introduced to capture the user's requirement that the meeting be organized by, respectively, booking a table at a nearby Italian restaurant $Book@Agt(S)@Obj(X)@Time(T)$ with constraints, $Near(X, H), Italian(X)$ and then telling Brian after it is reserved at time T_r (i.e., at time $T_t > T_r$). The system further decomposes this structure by adding sub-actions corresponding to $Id4$ and $Id5$: the choice of the restaurant Zingari (i.e., Z) is captured in the constraint $X = Z$ in $Id4$. Collectively, the DIS (minus the Id4 and Id5 boxes which have not been discussed and are planned in the background by the system) reflects the dialogue intentional structure.

3 Semantics of DISs

The semantics of any DIS in canonical form is specified by translating it into an FOL formula in a meta-language, $\mathcal{L}$. The translation of a DIS, $\mathcal{D}$, relative to a world, w, and an intention base, I, is written $\|\mathcal{D}\|^I_w$. The specification of the translation function makes use of a reification approach similar to that employed in the context of reasoning about knowledge (Moore 1985). The meta-language, $\mathcal{L}$, contains: (1) the usual logical connectives, $\{\wedge, \supset, \neg\}$, that stand for conjunction, implication and negation, respectively; (2) a set of meta-language constants that stand for variables and constants in the object language (i.e., the DIS language); and (3) a

[2] See Appendix for an explanation of the time index "T4".

1. $\|R(T_1,\ldots,T_n)\|^I_w = r(\|t_1\|^I_w \ldots, \|t_n\|^I_w)$, r a function
2. $\|T_1 = T_2\|^I_w = eq(t_1, t_2)$
3. $\|\langle V, C\rangle\|^I_w = exists(v, \|C\|^I_w)$,
4. $\|\neg\phi\|^I_w = not(\|\phi\|^I_w)$
5. $\|\phi \Rightarrow \psi\|^I_w =$
   $all(\{v_1,\ldots,v_m\}, \|C_1\|^I_w \& \ldots \& \|C_n\|^I_w \rightarrow \|\psi\|^I_w)$,
   where $\phi = \langle\{V_1,\ldots,V_m\}, \{C_1,\ldots,C_n\}\rangle$
6. $\|\langle\langle Id, A, V, T, Act, C, S\rangle\rangle\|^I_w =$
   $exists(\|vars^*(Id, I)\|^I_w, do(\alpha)\ \&\ \|cstr^*(Id, I)\|^I_w)$,
and $\alpha = act@id(id)@agt(a)@time(t)@tree(Id, I, w)$.
7. $\|Int[\langle V, T, C\rangle]\|^I_w = int(Holds(\|\langle V, C\rangle\|^I_w, t))$,
8. $\|\langle V, T, \mu\rangle\|^I_w = (\exists v_1 \ldots \exists v_n) holds(\|\mu\|^I_w, w, t)$,
   where $V = \{v_1,\ldots,v_n\}$.

The above make use of the following definitions:
$vars(Id, I) = V, \langle\langle Id, \_, V, \_, \_, \_, \_\rangle\rangle \in I$
$cstr(Id, I) = C, \langle\langle Id, \_, \_, \_, \_, C, \_\rangle\rangle \in I$
$vars^*(Id, I) = \bigcup_{s \in subs^*(Id,I)} vars(s, I)$
$cstr^*(Id, I) = \bigcup_{s \in subs^*(Id,I)} cstr(Id, I)$

Figure 2: Translation from DIS to FOL.

set of meta-language functions that stand for predicates and functions in the object language. In addition, $\mathcal{L}$ includes a single predicate symbol, $holds$, that ranges over terms, worlds and times: $holds(p, w, t)$. The term p can also have the complex form $int(Holds(q, t'))$—with upper-case term $Holds$. We use abstract syntax for logical operators in $\mathcal{L}$; thus, $holds(p\ \&\ q, w, t) \equiv (holds(p, w, t) \wedge holds(q, w, t))$, $holds(p \leftrightarrow q, w, t) \equiv (holds(p, w, t) \equiv holds(q, w, t))$, and $holds(not(p), w, t) \equiv \neg holds(p, w, t)$. In addition, if V is a set, $\{v_1,\ldots,v_n\}$, we write $exists(V, \phi)$ as shorthand for $exists(v_1, \ldots, exists(v_n, \phi)\ldots)$. To report that act-type α is performed by doing act-type β, we write: $do(\alpha@method(\beta))$.

Figure 2 gives the semantics of DISs. It assumes that all variables declared in (sub-)actions are unique as well as cross-world identity for constants, terms, predicate and function names. Possible worlds reflect alternative futures for intentions. We assume that there is a function, D, that takes a name in the object language and returns the corresponding name in the meta-language. These assumptions lead to the constraints $D(T) = t, D(P) = p$, etc.. Object-language elements will be in upper case and meta-language elements in lower case. We add the axioms:

$$holds(do(exists(v, do(e))), w, t) \equiv holds(exists(v, do(e)), w, t) \quad (1)$$

$$holds(do(not(x)), w, t) \equiv \neg holds(do(x), w, t) \quad (2)$$

$$holds(do(\alpha@time(t)), w, t') \equiv holds(do(\alpha), w, t) \quad (3)$$

We extend $\|.\|$ to any intention base, IS:

$$\|IS\|_w = \{\|I\|^I_w \mid I \in IS\}$$

We require that intentions be consistent: for any intention base, IS, it is not the case that both ϕ and $\neg\phi \in \|IS\|^I_{w,t}$. The semantics for intention is in FOL; we reify possible worlds, adopting modal logic System K (Chellas, 1980) where $acc_i(.,.,.,.)$ is a serial accessibility relation:

$$holds(int(a, Holds(p, t')), w, t) \equiv \forall w'.acc_i(a, w, w', t) \supset holds(p, w', t')$$

Here is an example of the FOL translation after utterance (5) (the θ_i's correspond to the act types in the Idi - see Appendix):

$$\begin{aligned} &holds(int(Holds(exists(\{t, x, t_t, t_r\} \\ &\quad do(\theta_1@id(id_1)@method(\theta_2@id(id_2) \\ &\quad @method(\theta_4@id(id_4))@method(\theta_5@id(id_5)) \\ &\qquad @method(\theta_3@id(id_3))) \\ &\quad \&\ restaurant(x)\ \&\ meeting(m) \\ &\quad \&\ near(x, h)\ \&\ italian(x)\ \&\ gt(t, t_3) \\ &\quad \&\ gt(t_r, t)\ \&\ gt(t_t, t_r)), t_p)), w_0, t_3). \end{aligned} \quad (4)$$

4 Intention revision

Most approaches to belief revision are founded on the idea of *minimal change*: to revise a set of beliefs, S, with some new p, where p is inconsistent with S, one should make the minimal change necessary to S to accommodate p. Our approach is *syntactic*, assigning greater significance to formulas, and their syntactic form, that appear in a *belief or intention* base (Nebel, 1989; Ortiz, 1999) than to the consequential closure of the corresponding base (the resulting *belief set*). Intention revision takes place in two steps within this framework as follows. Let S be an agent's current set of intentions. We translate S into its *predicate form* that explicitly refers to components of a DIS so that they can be modified according to the minimality criteria above. We use the same meta level language for constants and terms as in $\mathcal{L}$ above, augmented with special predicates to name the components of a DIS. If $\langle V_c, T_c, int[\langle V_t, T_t, \langle\langle Id_r, A_r, V_r, T_r, E_r, C_r, S_r\rangle\rangle\rangle]\rangle$ is a DIS, then its translation, for all

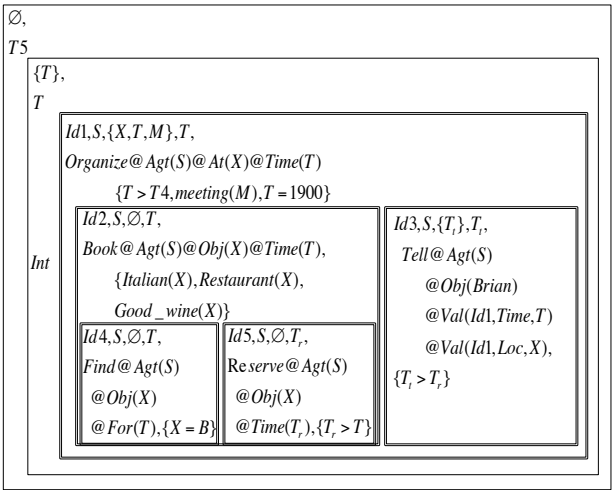


Figure 3: "And I'd like one with a good wine list."

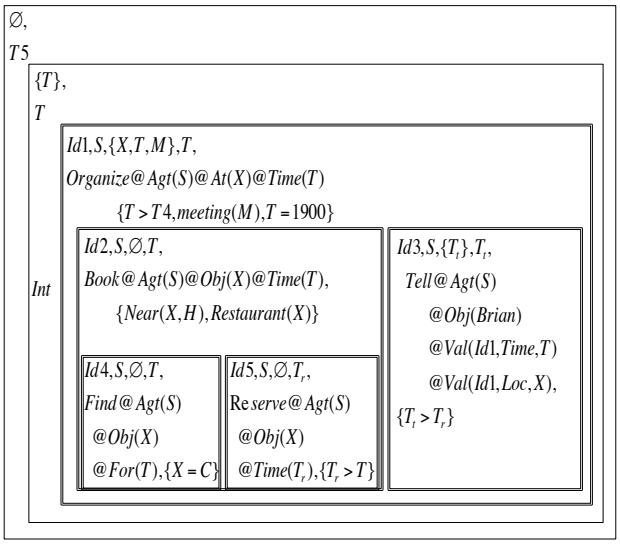


Figure 4: "Reserve a table at Chevy's instead."

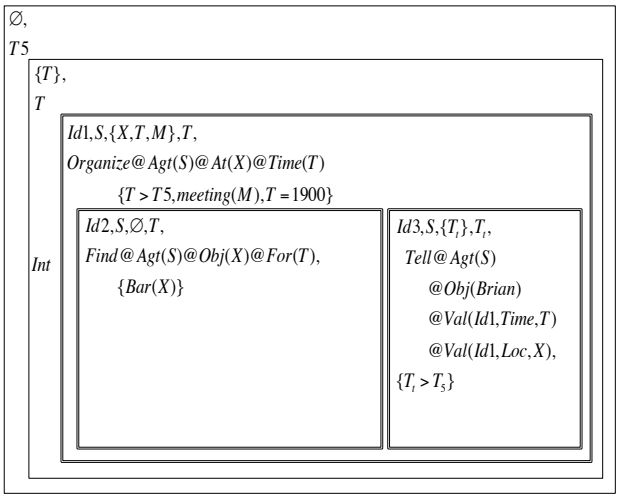


Figure 5: ". . . let's just go to a place for drinks."

$K_r \in C_r$, $U_r \in V_r, U_c \in V_c$, $B_r \in S_r$ is $\{id_r(id_r), agent_r(id_r, a_r), var_r(id_r, u_r), time_r(id_r, t_r), act_r(id_r, e_r), constr_r(id_r, k_r), sub_r(id_r, b_r), var_t(id_r, v_t), time_t(id_r, t_t), var_c(id_r, u_c), time_c(id_r, t_c)\} \cup \underline{nodes(ISB)}$; the latter is the set of predicate forms for each $\overline{S_r}$.

We call a collection of DISs plus associated plan nodes an *intention base* (IB). Given an intention base, ISB, we write $\underline{ISB}$ for the translation to predicate form and $\overline{ISB'}$ for the translation into canonical form of an intention base ISB' in predicate form.[3] Let S stand for an IB; to revise S with some ϕ we create a set of equivalence classes on S: $\{S_1, S_2, \ldots, S_n\}$ such that S_1 corresponds to those elements of S that are most important and S_n to those that are least important. To revise an intention base with some ϕ, we start with ϕ and add as much of each S_i that is consistent. Revisions involve either the addition or removal of (sub)actions or constraints from or to an IB.

The appendix provides the formal definition for intention revision and a derivation of the transformation between intention structures corresponding to some of the possible continuations of our target dialog. Here, we present the general idea using the box notation for canonical DISs. The purpose of partitioning the DIS boxes is to inform the revision process. A new utterance, u, is to be interpreted as is done in plan-based theories, as contributing somehow to the current intentional structure of the discourse, which in turn is closely related to the task structure or, in our case, to the DIS. To situate (the logical form of) u correctly in the current DIS (and, hence, the current discourse structure), we consider the contents of each box followed by each sub-box. If u corresponds to a new constraint, then we check for consistency first in the outer-most box. If it is consistent then we proceed to the sub-boxes. If it is not, we revise that sub-box with the new constraint. Similarly, if u corresponds to a new action and is consistent with the current box then we continue; if not, we delete that box and all of the sub-boxes which depended on it. These guidelines can be formalized to model the intention revision process using methods from the belief revision literature, as long as we operate on the DIS predicate form.

Figure 3 illustrates the transformation that takes place after utterance (8a). The formula

[3] See (Ortiz and Hunsberger, 2013) for details.

$good\_wine(X)$ is consistent with the contents of $Id2$. However, box $Id4$ and $Id2$ are not jointly consistent with the constraint $X = Z$ (shown in Figure 1). The maximal subset that is consistent with the new constraint is one that contains all of the existing constraints except $X = Z$ and $near(x, h)$. That is therefore deleted (assuming $near(x, h)$ has lowest priority) and replaced with the new constraint. Consistency is checked on the FOL translation (4).

Continuation (8b), results in the DIS shown in Figure 4. The new constraint is not consistent with the constraint in $Id4$ related to the wine; hence, the latter is deleted. Next, the new constraint is not consistent with the choice $X = B$ in Figure 1; hence, it is replaced by the new one, $X = C$. As a final example, (8d), is inconsistent, by appeal to pragmatic world knowledge - one doesn't go to restaurants just for drinks – with all of $Id2$. In the formal definition of $\underline{ISB}$, if a negated action is in ISB then the box (and sub-boxes) corresponding to it is deleted. Hence, $Id2$ and all of its sub-boxes are deleted and the new, $bar(X)$ constraint is added (Figure 5).[4] Note that, in all of these possible continuations, the side-effects to the $Id3$ component does not have to be modified. The desired changes result simply because of the way that the intention is structured and the locality of variables.

5 Implementation

We are developing a collaborative dialogue manager (CDM) that embodies the ideas described in this paper. We are testing it in a living room setting where the user asks a TV equipped with speech recognition software and natural language (NL) understanding for help in, for example, locating, playing or recording entertainment available from different content providers. CDM is an extension of Disco (Rich and Sidner, 2012), an open source dialogue development framework based on Collaborative Discourse Theory (Grosz and Sidner, 1986; Grosz and Kraus, 1996; Lochbaum, 1998). It views a VPA dialogue as a process of plan augmentation, where the purpose of the dialogue is for the system and the user to collaborate on a complete SharedPlan to meet a user's inferred intention. Each user utterance is processed by an NL pipeline consisting of named entity recognition (NER) followed by morphological, syntactic and semantic processing. The CDM then initiates a planning process by first accessing a recipe library consisting of, essentially, a collection of hierarcical networks (HTNs) that decompose high-level task (goal) structures; the recipes are written in the the ANSI/CEA-2018 standard (Rich, 2009). If a plan cannot be constructed, then one of several builtin utterance generation rules is fired and a system utterance is generated in order to acquire the necessary information from a user to further the planning process. The cycle continues until a complete plan is formed for the user's intention.

We have extended CDM with the DIS framework. CDM provides the procedural, stack-based management of *attentional state* (the "in-focus" portion of the DIS) and the dialogue segmentation. CDM generates either group-level (an intention that a group — e.g., the system and the user — perform some group action) or individual-level DISs (Hunsberger and Ortiz, 2008). The DISs depicted below include fields for two types of variables: *ExVars* and *DefVars*, those that the intending agent is free to assign values to and those determined by some other agent, respectively. There are three points where a DIS may be generated or updated: when a user utterance is interpreted, during the plan generation and decomposition process, or when a system utterance is generated. The plan augmentation process makes use of DISs directly. We will use the following simple dialogue (Figure 6) to illustrate the operation of the CDM. The "boxes" in the figures are added for readability only; the system only currently produces ASCII text with the explicit references to sub-boxes shown in the figures.

1. User: Play a James Bond movie without Sean Connery
2. System: Ok. Which one would you like to see? Skyfall or Tomorrow Never Dies?
3. User: Skyfall.

Figure 6: The James Bond example

At the start of the dialogue, CDM generates a group intention for the system to display a movie, m, for the user to watch that meets the constraints, i.e., a James Bond movie without Sean Connery. (The Group DIS, idG, is shown in Figure 7).

The system then identifies a relevant recipe from its recipe library. The top level recipe *Com-*

[4]Note: in the actual implementation the user is *asked* before deleting the $near(x, h)$ constraint because the meeting node has a location property as well.

```
ID/Agt/Grp: idG/g/GROUP
ExVars: m, s, j
ActType: Display@agt(System)@obj(m)
Conds: VideoConceptualWork(m),
CharacterinConceptualWork(j, m),
~VideoConceptualWorkActors(m, s), NarrativeRole(j),
Name(s, "Sean Connery"), Name(j, "James Bond"),
Person(s), Male(s)
```

Figure 7: DIS generated after user utterance 1

mandPlayVCW (Figure 8) is chosen to fulfill the group intention; the system adopts it as its own intention. Consequently, a new DIS for the system idS is generated and the group DIS idG is updated to have idS as its subBox (Figure 9). During plan decomposition, CDM generates two subBoxes $idSLookup$ and $idSPlay$ for the system DIS idS to reflect the decomposed steps (*Lookup* and *PlayVCW*) in the *CommandPlayVCW* recipe (Figure 10).

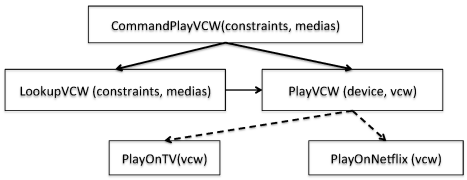


Figure 8: Recipe CommandPlayVCW

```
ID/Agt/Grp: idG/g/GROUP
ExVars: s, j
DefVars: (m, {Do(CommandPlayVCW@agt(System)@obj(_), …},
System, idS, ms)
ActType: Display@agt(System)@obj(m)
Conds: …
SubBoxes: idS

ID/Agt: idS/System
ExVars: ms
ActType: CommandPlayVCW@obj(ms)
Conds: VideoConceptualWork(ms),
CharacterinConceptualWork(j, ms),
~VideoConceptualWorkActors(ms, s)
```

Figure 9: DIS's updated user utterance 1

The *Lookup* method returns two movies *Casino Royale* and *A View to a Kill*, and the *Ask.Which* utterance generation rule is fired, producing the system utterance 2 of Figure 6. The CDM updates the system's DISs and generates a new DIS idU for the user to reflect that the movie, ms, to be played depends on the user's selection (Figure 11).

When the user chooses *Skyfall*, the system interprets that as a response to the *Ask.Which* question, and integrates the new user contribution into the user DIS *idU* by updating the ExVar mu to DefVar (Figure 12). This plan refining process continues

```
ID/Agt: idS/System
ExVars: ms
DefVars: (msSet, {Do(Lookup@results(_), …}, System,
idSLookup, msLookupSet)
ActType: CommandPlayVCW@obj(ms)
Conds: FORALL ms IN msSet, …
SubBoxes: idSLookup, idSPlay

ID/Agt: idSLookup/System
ExVars: msLookupSet
ActType: Lookup@results(msLookupSet)
Conds: FORALL msLookup IN msLookupSet,
VideoConceptualWork(msLookup),
CharacterinConceptualWork(j, msLookup),
~VideoConceptualWorkActors(msLookup, s)


ID/Agt: idSPlay/System
ActType:PlayVCW@obj(ms)
```

Figure 10: DIS's updated after plan decomposition

```
ID/Agt: idS/System
DefVars: (ms, {Sel(User,_,"ms", idS),_ IN msSet},User, idU,mu),
         (msSet, {Do(Lookup@results(_), …}, System, idSLookup,
             msLookupSet)
ActType: CommandPlayVCW@obj(ms)
SubBoxes: idSLookup, idSPlay, idU

ID/Agt: idSLookup/System
ExVars: msLookupSet
ActType: Lookup@results(msLookupSet)
Conds: FORALL msLookup IN msLookupSet,
VideoConceptualWork(msLookup),
CharacterinConceptualWork(j, msLookup),
~VideoConceptualWorkActors(msLookup, s)

ID/Agt: idSPlay/System
ActType:PlayVCW@obj(ms)

ID/Agt: idU/User
ExVars: mu
Conds: Sel(User, mu, "ms", idS),
             mu IN msSet
```

Figure 11: DIS's updated after CDM generates system utterance 2

until a complete plan is formed.

In CDM, the utterance interpretation and generation processes as well as the plan decomposition and refinement processes are grounded in the DIS framework. If we continue the dialogue of Figure 6 with the user utterance "How about one with Ursula Andress," the current CDM requires a separate external consistency check; we are currently integrating that process into the system.

6 Summary and related work

We have focussed on re-planning dialogues common to scenarios involving a user and a personal assistant. In such settings (in contrast to, say, master-apprentice dialogues) a user often does not know whether a set of proposed constraints on a task will be satisfiable: new suggested constraints can conflict with already expressed ones. Further, the process of discourse segmentation can interact in a negative way with the constraint relaxation process, resulting in incorrectly situating a new utterance within a segmented discourse.

```
ID/Agt: idS/System
DefVars: (ms, {Sel(User,_,"ms",idS), _ IN msSet},User,idU,mu),
         (msSet, {Do(Lookup@results(_), …}, System, idSLookup,
              msLookupSet)
ActType: CommandPlayVCW@obj(ms)
SubBoxes: IdSLookup

 ID/Agt: idSLookup/System
 ExVars: msLookupSet
 ActType: Lookup@results(msLookupSet)
 Conds:  FORALL msLookup IN msLookupSet,
 VideoConceptualWork(msLookup),
 CharacterinConceptualWork(j, msLookup),
 ~VideoConceptualWorkActors(msLookup, s)

 ID/Agt: idSPlay/System
 ActType:PlayVCW@obj(ms)

 ID/Agt: idU/User
 DefVars: (mu, Skyfall)
 Conds: Sel(User,mu,"ms",idS),
              mu IN msSet
```

Figure 12: DIS's updated after user utterance 2

Such dialogues require some non-monotonic form of intention revision during the process of accommodating a new utterance into the existing dialog. We applied the DIS framework developed to model intention revision in rational agents. Intentions are structured to inform an incremental revision process: rather than completely eliminating any conflicting intention, the approach first attempts to minimally revise the contents of an individual intention; in the process, side-effects are automatically handled. Since DISs are based on a dynamic logic approach similar to DRT, a bridge is created between plan-based dialogue approaches and rigorous accounts to discourse meaning found in DRT and argued to be missing from cognitive approaches (Asher and Lascarides, 2003).

Segmented Discourse Representation Structures (SDRS) structure discourses using discourse relations; however, the rich and revisable hierarchical intention structures that we have argued for are absent. Neither DRT nor SDRS deal with the revision of structures in the case of inconsistencies. Recent work has examined the modification of decision theoretic agent preferences during dialogue (Cadilhac et al., 2011). However, their plan-correction methods do not deal with side-effects nor are they tightly linked to a formal representation of intentions. In addition, desires in the theory of SharedPlans, on which we are basing our work, formalizes desires instead as potential-intentions-to perform some action. The Collagen system maintained a segmented history of the dialogue which a user could manually examine and manipulate (Rich and Sidner, 1998): a user could retract, say, an action in a recipe plan tree and a truth maintenance system would then retract logical dependencies. Our system instead performs such "undos" automatically.

Work on correction and denials that retracts contextual information appearing earlier in a discourse is related (van Leusen, 2004; Maier and van der Sandt, 2003). That work differs, however, in that corrections and denials are explicit and discourses are not structured into larger segments. Work in SDRS in this area has not dealt with the problem of revision (Lascarides and Asher, 2009). Finally, user-initiated correction dialogs (Lochbaum, 1998)) are somewhat different as they are triggered by an observed plan obstacle.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

A. Cadilhac, N. Asher, F. Benamara, and A. Lascarides. 2011. Commitments to preferences in dialogue. In *Meeting of the SIG on Discourse and Dialogue*.

Brian F. Chellas. 1980. *Modal Logic: An Introduction*. Cambridge University Press.

Barbara J. Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86(1):269–357.

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

Luke Hunsberger and Charles Ortiz. 2008. Dynamic Intention Structures I: A theory of intention representation. *Autonomous Agents and Multiagent Systems*, pages 298–326.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments. *Journal of Semantics*, 26(2):109–158.

Karen E. Lochbaum. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 34(4):525–572.

E. Maier and R. van der Sandt. 2003. Denial and correction in layered DRT. In *Proceedings of diaBruck*.

Robert C. Moore. 1985. A formal theory of knowledge and action. In *Formal Theories of the Commonsense World*. Ablex Publishing Corporation.

Bernhard Nebel. 1989. A knowledge level analysis of belief revision. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 301–311.

Charles L. Ortiz and Luke Hunsberger. 2013. On the revision of dynamic intention structures. In *Eleventh International Symposium on Logical Formalizations of Commonsense Reasoning*.

Charles L. Ortiz. 1999. Explanatory update theory: Applications of counterfactual reasoning to causation. *Artificial Intelligence*, 108:125–178.

Charles Rich and Candace L. Sidner. 1998. COLLAGEN: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction*, 8(3/4):315–350.

Charles Rich and Candace Sidner. 2012. Using collaborative discourse theory to partially automated dialogue tree authoring. In *14th International Conference on Intelligent Virtual Agents*.

Charles Rich. 2009. Building task-based user interfaces with ANSI/CEA-2018. *IEEE Computer*, 43(8):20–27.

Noor van Leusen. 2004. Incompatibility in context: a diagnosis of correction. *Journal of Semantics*, 21(4):415–415.

Appendix: Worked out formal example

Definition 1 (Intention revision) *Let I and I' be DISs in predicate form and let S_i be the set of induced equivalence classes on I, $i \geq 1$. The prioritized removal of elements of I that conflict with $\neg\|I'\|$, which we write as $I \bullet I'$, is (Nebel, 1989):*

$$I \bullet I' = \{Y \subseteq I \mid \|\overline{Y}\| \not\vdash \neg\|\overline{I'}\|, \; Y = \cup_i Y_i, i \geq 1 \; \forall i \geq 1 : (Y_i \subseteq S_i, \; \forall X : Y_i \subset X \subseteq S_i \rightarrow (\bigcup_{j=1}^{i-1} Y_j \cup X) \vdash \neg\|\overline{I'}\|)\}$$

We can define the operation of intention revision by some I' that is inconsistent with I as:

$$I \star I' = \cap_{(Y \in I \bullet I')} \cup I'$$

Starting with I' is first augmented with the maximal subset of S_1 that is consistent (via the translation to FOL). This is repeated for each maximal subset of the next equivalence class until no additional elements of S can be consistently added.

We consider five steps, at times $t_1 < t_2 < \ldots < t_5$, of intention formation. For any t_i, the canonical form of the IB is $IS(t_i)$ and $IS[t'/t]$ indicates that all instances of t in IS are substituted by t'.

Step 1. The system (s) intends at time t_1 to organize a meeting later at $t = 1900$ (7pm). (We collapse utterances (1-3) in this step.)

$$IS(t_1) = \{\langle\emptyset, T_1, Int[\langle\{T\}, T, \langle\langle Id_1, S, \{X, M, T\}, T, \Theta_1, \{T > T_1, Meeting(M), T = 1900\}, \emptyset\rangle\rangle]\rangle\}$$

s.t., $\Theta_1 = Organize@Agt(S)@At(X)@Time(T)$.

The predicate form of this intention is:

$$\underline{IS(t_1)} = \{\, id_r(id_1), agt_r(id_1, s), var_r(id_1, m), time_c(id_1, t_1), time_t(id_1, t), time_r(id_1, t), var(id_1, x), act_r(id_1, \theta_1), var_t(id_1, t), constr_r(id_1, gt(t, t_1)), constr_r(id_1, eq(t, 1900)) \; constr_r(id_1, meeting(m)) \,\}$$

s.t., $\theta_1 = organize@agt(s)@at(x)@time(t)$

and $gt(t, t_1)$ is the metalanguage form of $T > T_1$.

The FOL form, relative to the real world, w_0, is:

$$\|IS(t_1)\|_{w_0} = holds(int(Holds(exists(\{t, x, m\}, do(\theta_1@id(id_1)) \;\&\; gt(t, t_1) \;\&\; eq(t, 1900) \;\&\; meeting(m)), t)), w_0, t_1)$$

Step 2. The meeting is organized by booking a restaurant near the hotel (H) and telling Brian. The "tell" action sends parameter values to Brian:

$$\underline{IS(t_2)} = \underline{IS(t_1)}[t_2/t_1] \star \{id(id_2), id(id_3), sub_r(id_1, id_2), agt(id_2, s), time(id_2, t), sub_r(id_1, id_3), agt(id_3, s), time(id_3, t_t), var(id_3, t_t), act(id_2, \theta_2), act(id_3, \theta_3), constr(id_2, restaurant(x)), constr(id_2, near(x, h)), constr(id_2, italian(x), constr(id_3, gt(t_t, t))\}$$

where $\theta_2 = book@agt(s)@obj(x)@time(t)$

and $\theta_3 = tell@agt(s)@obj(brian)@val(id_1, time, t) @val(id_1, loc, x)$

In canonical form we have:

$$IS(t_2) = \{\langle\emptyset, T_2, Int[\langle\{T\}, T, \langle\langle Id_1, S, \{X, M, T\}, T, \Theta_1, \{T > T_2, Meeting(M), T = 1900\}, \{Id_2, Id_3\}\rangle\rangle]\rangle, \langle\langle Id_2, S, \emptyset, T, \Theta_2, \{Restaurant(X), Near(X, H), Italian(X)\}, \emptyset\rangle\rangle, \langle\langle Id_3, S, \{T_t\}, T_t, \Theta_3, \{T_t > T\}, \emptyset\rangle\rangle\}$$

where $\Theta_2 = Book@Agt(S)@Obj(X)@Time(T)$

and: $\Theta_3 = Tell@Agt(S)@Obj(Brian) @Val(Id_1, Time, T)@Val(Id_1, Loc, X)$.

Step 3. The system decides to book the table by finding and reserving a restaurant.

$$\underline{IS(t_3)} = \underline{IS(t_2)}[t_3/t_2] \; \star \; \{id(id_4), id(id_5), sub(id_2, id_4), sub(id_2, id_5), agt(id_4, s), agt(id_5, s), time(id_4, t), time(id_5, t_r), act(id_4, \theta_4), act(id_5, \theta_5), constr(id_5, gt(t_r, t))\}$$

where $\theta_4 = find@agt(s)@obj(x)@for(t)$

and $\theta_5 = reserve@agt(s)@obj(x)@time(t_r)$.

In canonical form, the result is:

$$
\begin{aligned}
IS(t_3) = \{&\langle \emptyset, T_3, Int[\langle \{T\}, T,\\
&\langle\!\langle Id_1, S, \{X, M, T\}, T, \Theta_1, \{T > T3,\\
&Meeting(M), T = 1900\}, \{Id_2, Id_3\}\rangle\!\rangle]\rangle,\\
&\langle\!\langle Id_2, S, \emptyset, T, \Theta_2, \{Restaurant(X),\\
&Near(X, H), Italian(X)\}, \{Id_4, Id_5\}\rangle\!\rangle,\\
&\langle\!\langle Id_3, S, \{T_t\}, T_t, \Theta_3, \{T_t > T, T > T_2\}, \emptyset\rangle\!\rangle,\\
&\langle\!\langle Id_4, S, \emptyset, T, \Theta_4, \emptyset, \emptyset\rangle\!\rangle,\\
&\langle\!\langle Id_5, S, \emptyset, T_r, \Theta_5, \{T_r > T\}, \emptyset\}\\
\text{s.t., } \Theta_4 &= Find@Agt(S)@Obj(X)@For(T),\\
\Theta_5 &= Reserve@Agt(S)@Obj(X)@Time(T_r)
\end{aligned}
$$

The FOL translation is given by formula (4).

Step 4. The user selects Zingari (Z) and the system adds it to the intention structure.

$$\underline{IS(t_4)} = \underline{IS(t_3)}[t_4/t_3] \star constr(id_4, eq(x, z))$$

The canonical form is given in Figure 1. It follows that the system also intends to tell Brian of the location, Zingari, and to reserve a table there.

Step 5. The system revises its intention to include the constraint of a good wine list. (Previously, "$\star$" corresponded to set union) We revise the intention and assume a joint user-system selection of Barbacco ($x = b$). The knowledge base also contains, with highest priority, the following:

$$
\begin{aligned}
holds(good\_wine(y) \leftrightarrow eq(y, b), w, t) \quad (5)\\
holds(near(z, h)\ \&\ \sim near(b, h), w, t)\\
holds(italian(z)\ \&\ italian(b), w, t)\\
holds(restaurant(z)\ \&\ restaurant(b)w, t)
\end{aligned}
$$

We have,

$$
\begin{aligned}
&\underline{IS(t_5)} =\\
&\quad IS(t_4)[t_5/t_4] \star \{constr(id_4, good\_wine(x))\}
\end{aligned}
$$

The following are the set of priority classes that we will use, separating the tree, constraints and assignments. General or more detailed rules can be written (Ortiz and Hunsberger, 2013).

$$
\begin{aligned}
S_1(IS(t_4)) = \{&id_r(id_1), agt_r(id_1, s), var_r(id_1, m),\\
&var_r(id_1, x), var_t(id_1, t), act_r(id_1, \theta_1),\\
&time_c(id_1, t_4), time_t(id_1, t), time_r(id_1, t),\\
&id(id_2), sub_r(id_1, id_2), agt(id_2, s), sub(id_2, id_5),\\
&id(id_3), sub_r(id_1, id_3), agt(id_3, s), id(id_5),\\
&id(id_4), sub(id_2, id_4), agt(id_4, s), agt(id_5, s),\\
&time(id_2, t), time(id_3, t_t), time(id_4, t), time(id_5, t_r),\\
&act(id_2, \theta_2), act(id_3, \theta_3), act(id_4, \theta_4),\\
&var(id_3, t_t), act(id_5, \theta_5)\}\\
S_2(IS(t_4)) = \{&constr(id_3, gt(t_t, t_r)),\\
&constr(id_5, gt(t_r, t))\\
&constr(id_2, restaurant(x)), constr(id_2, italian(x)),\\
&constr_r(id_1, gt(t, t_4)), constr_r(id_1, meeting(x)),\\
&constr(id_3, gt(t_t, t)), constr(id_5, gt(t_r, t))\},\\
S_3(IS(t_4)) = \{&constr(id_4, eq(x, z)),\\
&constr(id_1, eq(t = 1900))\}\\
S_4(IS(t_4)) = \{&constr(id_2, near(x, h))\}
\end{aligned}
$$

S_1 and S_2 go through but $constr(id_4, eq(x, z))$ (S_3) and S_4 conflict and are not included in $\underline{IS(t_5)}$. To see this, we translate to FOL, and apply axiom (1):

$$
\begin{aligned}
&\|IS(t_4)\|_{w_0} =\\
&\quad holds(int(Holds(exists(\{t, x, t_t, t_r, m\}\\
&\quad do(\theta_1@id(id_1)@method(\theta_2@id(id_2)\\
&\quad @method(\theta_4@id(id_4))@method(\theta_5@id(id_5))\\
&\qquad @method(\theta_3@id(id_3)))\\
&\quad \&\ restaurant(x)\ \&\ gt(t_t, t)\ \&\ gt(t, t_3)\\
&\quad \&\ italian(x)\ \&\ near(x, h)\ \&\ eq(t, 1900)\\
&\quad \&\ gt(t_r, t)\ \&\ gt(t_t, t_r)\ \&\ good\_wine(x)\\
&\quad \&\ meeting(m)\ \&\ eq(x, z))))), w_0, t_4).
\end{aligned}
$$

We eliminate *holds* expressions by referring to the accessibility relation and (1), converting "&" to conjunction. The result is inconsistent, given axioms (5). Similarly, S_4, is also inconsistent. Barbacco can now be inserted into the DIS. The result (shown in Figure 3). It follows that the system will tell Brian that the location is Barbacco, as desired.

The remaining cases are handled similarly. In choosing a bar, an axiom would preclude that together with booking a restaurant; by S_1 and the mapping back to canonical form, we would have an inconsistency, retracting the entire "box" for $Id2$.

Revealing Resources in Strategic Contexts

Jérémy Perret
IRIT,
Univ. Toulouse, France

Stergos Afantenos
IRIT,
Univ. Toulouse, France

Nicholas Asher
IRIT, CNRS,
France

Alex Lascarides
School of Informatics
Univ. Edinburgh, UK

{perret,stergos.afantenos,asher}@irit.fr
alex@inf.ed.ac.uk

Abstract

Identifying an optimal game strategy often involves estimating the strategies of other agents, which in turn depends on hidden parts of the game state. In this paper we focus on the win-lose game *The Settlers of Catan* (or *Settlers*), in which players negotiate over limited resources. More precisely, our goal is to map each player's utterances in such negotiations to a model of which resources they currently possess, or don't possess. Our approach comprises three subtasks: (a) identify whether a given utterance (dialogue turn) reveals possession of a resource, or not; (b) determine the type of resource; and (c) determine the exact interval representing the quantity involved. This information can be exploited by a *Settlers* playing agent to identify his optimal strategy for winning.

1 Introduction

When resources are limited, there is a fine line between agents cooperating and competing with one another for those resources, especially in a win-lose game. The goal of every rational agent is to maximize his *expected utilities* by finding *equilibrium strategies*: that is, an action sequence for each player that is optimal in that no player would unilaterally deviate from his action sequence, assuming that all the other players perform the actions specified for them (Yoam Sholam and Kevin Leyton-Brown, 2009). Calculating equilibrium strategies thus involves reasoning about what's optimal for the other players, which in turn depends on which resources they possess and which resources they need. However, almost every kind of bargaining game occurs in a context of imperfect information (Osborne and Rubinstein, 1994), where the opponent's current resources are hidden or non-observable.

Indeed, imperfect information often results from deliberate obfuscation: if an opponent can accurately identify your resources then they can exploit it for their own strategic advantage. For instance, in *The Settlers of Catan* (or *Settlers*), our chosen domain of investigation here, Guhe and Lascarides (2014) develop a *Settlers* playing agent where game simulations show that making the agent omniscient about his opponents' resources enables him to achieve more successful negotiations (i.e., a significantly higher proportion of his trade offers are accepted) and a significantly higher win rate than his non-omniscient counterparts. So it is rational for players to balance achieving their desired trades with revealing as little as possible about their own resources, while at the same time attempting to elicit information about their opponents' resources.

In negotiations using natural language dialogue, eliciting information about an opponent's resources is often realized as a question; the opponent, on realizing the question's purpose, often avoids revealing their resources in their response. They use various communicative strategies to achieve this effect, such as making a counteroffer, being vague, or simply changing the subject.

In this paper, we are interested in determining how players can extract information about an opponent's resources from what they say during negotiation dialogues. In order to study how people commit, or don't commit, to the resources they have (or don't have), we have used a corpus of negotiation dialogues that take place during the win-

lose game *The Settlers of Catan* in order to learn a statistical model that maps the utterances of the players to their commitments concerning the kind and number of resources they possess. In section 2 we describe our corpus in detail, as well as the phenomena that we are trying to capture. In section 3 we describe the annotation procedure that we have followed in order to obtain training and testing datasets. Section 4 describes the experiments we have performed and the results we have obtained. Section 5 describes the related work and conclusions and future work are in section 6.

2 The Corpus

Our model is trained on an existing corpus (see Afantenos et al. (2012)) of humans playing an online version of the game *The Settlers of Catan* (or *Settlers*, Teuber (1995); `www.catan.com`). *Settlers* is a win-lose game board game for 2 to 4 players. Each player acquires resources (ore, wood, wheat, clay, sheep) and uses them to build roads, settlements and cities. This earns Victory Points (VPs); the first player with 10 VPs wins. Players can acquire resources via the dice roll that starts each turn and through trading with other players—so players converse to negotiate trades. 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. Players can also lose resources: a player who rolls a 7 can rob from another player and any player with more than 7 resources must discard half of them. What's robbed or discarded is hidden, so players lack complete information about their opponents' resources. Consequently, agents can, and frequently do, engage in 'futile' negotiations that result in no trade (i.e., they miscalculate the equilibria).

Players in the corpus described in Afantenos et al. (2012) must chat in an online interface in order to negotiate trades, and each move in the chat interface is automatically aligned with the current game state—so one can compare what an utterance reveals about possessed resources with what the speaker actually possesses, and so identify examples of obfuscation (e.g., see Table 1). The corpus consists of 59 games, and each game contains dozens of individual negotiation dialogues, each dialogue consisting of anywhere from 1 to over 30 dialogue turns. In our experiments, we have used 7 games consisting of more than 2000 dialogue turns (see Section 3).

Table 1 contains an excerpt from one of the dialogues. In turn 157 the player "gotwood4sheep" asks if anyone has any wood, implying that he wants to negotiate an exchange of resources where he receives wood. Player "ljaybrad123" is the first to reply, negatively, implicating that he has no wood. Turn 158 is thus annotated with the information that the player "ljaybrad123" is revealing that he has 0 wood.[1] In turn 159 player "gotwood4sheep" persists in his attempt to negotiate, referring directly to player "tomas.kostan" and making a more specific trade offer, of ore in exchange for wood. He has thus revealed that he possesses at least one ore. The player "tomas.kostan" acknowledges that he has wood (so this turn is annotated with the information that "tomas.kostan" has at least one wood) but that this resource is important to him. "tomas.kostan" then proposes 2 ore in exchange for 1 wood (again, this turn is annotated with the information that "tomas.kostan" possesses at least one wood). "gotwood4sheep" in turn 162 explicitly says that he has only one ore and not two, so this turn is annotated with the information that player "gotwood4sheep" has exactly 1 ore. In the end the negotiation fails since for "tomas.kostan" a wood is currently worth more to him than what "gotwood4sheep" is currently offering.

Note that revealed resources depend not only on the content of the individual utterance but also on its semantic connection to the discourse context. For example, the dialogue turn 158 (*no*) reveals nothing about resources on its own; it is the fact that it is connected to the question 157 with a QAP (Question-Answer-Pair) relation that commits "ljaybrad123" to having 0 wood. Similarly, 160 is an Acknowledgment to 159 and so reveals that "tomas.kostan" possesses at least one wood.

3 Annotations

The corpus has been annotated with information at multiple levels, including dialogue boundaries, turns within dialogues, speech acts (offers, counteroffers, refusals, etc.), as well as discourse relations following SDRT (Asher and Lascarides, 2003). Full details are in Afantenos et al. (2012); here we provide in Tables 2 and 4 statistics of only

[1] In this paper, we simplify our task by ignoring the fact that players can lie. As matter of fact, manual analysis of the corpus logs show that players rarely lie concerning their resources, preferring instead to conceal relevant information by avoiding giving a direct answer.

Dialogue turn	Player	Utterance
157	gotwood4sheep	anyone got wood?
158	ljaybrad123	no
159	gotwood4sheep	ore for a wood, tomas?
160	tomas.kostan	yes but i need mine
161	gotwood4sheep	ore more?
162	tomas.kostan	2 ore for a wood?
163	gotwood4sheep	i don't have 2, sorry, just the one
164	gotwood4sheep	early doors, early offers :)
165	tomas.kostan	then i cannot make you a deal
166	tomas.kostan	sry
167	gotwood4sheep	ah dommage :(

Table 1: Excerpt from a dialogue

Number of speech turns in dialogue	Dialogue count
1-5	112
6-10	63
11-15	23
16-20	13
21 and more	23

Table 2: Dialogue statistics

Data counts	
Number of games	7
Speech turns	2460
Relation count by type	
Question-answer pair	687
Comment	443
Continuation	250
Acknowledgement	230
Result	182
Q-Elab	161
Elaboration	150
Contrast	140
Explanation	79
Clarification question	52
Narration	43
Alternation	42
Correction	41
Parallel	40
Conditional	32
Background	19

Table 3: Annotation of discourse relations

the relevant annotations that we used to train our models. Our model mostly exploits the QAP and Q-Elab relations to infer revealed resources; see Section 4 for details, including the performance of our trained model for identifying discourse relations.

We manually annotated each utterance with its corresponding revealed resource. Two of this paper's authors were involved in this annotation effort. After a thorough examination of the dialogues in an initial game, they settled on the format of the annotations and the guide for performing the annotation task. The annotation format is as follows. Each speech turn corresponding to a revealed resource is annotated with a pair: a resource name, and the quantity interval which the player reveals, representing the lower and upper bound of the resource. For example, in dialogue turn 158 of table 1 player "ljaybrad123" declares that he has no wood, so this dialogue turn is annotated as **`(wood, [0,0])`**. In dialogue turn 159 player "gotwood4sheep" reveals he has at least one ore, so this turn is annotated as **`(ore, [1,+∞])`**. Revelations of multiple resources are associated with multiple pairs.

To test the consistency and difficulty of the task, both annotators independently annotated a single game after settling on the above format and instructions for annotation. Over 422 speech turns, the resulting kappa coefficient of inter-annotator agreement is **0.94**, enough to validate our annotation method. The remaining 6 games were then

Speech turns	2201
Dialogues	263
Word count	9121
Turns revealing resources	452 (21% of turns)

Table 4: Dataset overview

annotated, for which statistics can be found in tables 4 and 2. Most dialogues appear to be short, frequently consisting of comments on the game status, which do not call for answers. Trade negotiations are usually longer, with player emitting offers and counteroffers, sometimes competitively. Revelations of resources are present in 21% of dialogue turns.

4 Experiments and Results

4.1 Formulating the problem

As mentioned earlier, our goal is to predict whether a given turn reveals that its emitter possess a resource, and if so the type of the resource and its quantity in the form of an interval. Although players could potentially reveal having a specific number of resources (e.g., line 163 in table-1), in most cases the players reveal either having zero resources (interval $[0, 0]$) or having at least one (interval $[1, \infty]$), and in few occasions, players reveal that they have more than one (interval $[2, \infty]$) or exactly two resources ($[2, 2]$). In most of the cases, a revelation of having zero resources is manifested through the player rejecting a trade offer by stating that they don't have the resource desired by their opponent.

Using a single classifier to predict from an NL string the revelation of a particular type of resource, or no revelation of any resource, would involve classifying each utterance into 6 classes: one for each of the 5 types of resources, and one for revealing that no resources are possessed. But such a model would fail to take full advantage of the following facts. First, the NL strings that reveal a resource are relatively invariant, save for the particular resource type; in other words, the ways in which people talk about their possession of clay is the same as their talk about possessing wood, save for the words "clay" vs. "wood". Secondly, it is easy to specify the properties of a revelation (both the type of resource and quantity) when we know a given utterance exhibits a revelation. Given these observations, we decided to divide the prediction process into two subtasks:

- Determine if a given speech turn reveals a resource or not;
- For those utterances that do reveal a possessed resource, determine the type of resource and its associated quantity interval.

4.2 Features

Our goal was to learn a function

$$f : \mathcal{X} \mapsto \{0, 1\}$$

where every $\mathbf{x} \in \mathcal{X}$ corresponds to a vector representing a dialogue turn and $\{0, 1\}$ represents the fact that there is a revelation concerning an underspecified resource from the part of the dialogue act emitter.

The features that we have extracted for every dialogue turn can be summarized in the following categories:

- Contextual features: positioning of the turn in the dialogue;
- Lexical features: single words present in the utterance;
- Pattern-related features: recurring speech structures associated with revealed resources;
- Relational features: discourse relationships with other turns.

These features are listed more extensively in Table 5. Non-relational features are extracted directly from the underlying text. In order to compute the relational features—essentially whether a pair of dialogue turns are linked with a *Question-answer pair* (QAP) or a *Question-Elaboration* (Q-Elab) discourse relation—we used the results of a separate classifier for the prediction of discourse relations. This classifier was trained on 7 games consisting of 2460 dialogue turns. We used a MaxEnt model, as in the case of predicting revealed resources (see below for more details). We selected, for this classifier, a subset of the feature set used for the task of predicting revealed resources. More specifically, we used only the *Contextual* and *Lexical* features shown in Table 5. Although the model we have used was a general one, capable of predicting the full set of discourse relations listed in Table 3, for this series of experiments we were only interested in the QAP and Q-Elab relations. Results for these relations are shown in Table 6.

Category	Description
Contextual	Speaker initiated the dialogue
Contextual	First utterance of the speaker in the dialogue
Contextual	Position in dialogue
Lexical	Contains resource name
Lexical	Ends with exclamation mark
Lexical	Ends with interrogation mark
Lexical	Contains possessive pronouns
Lexical	Contains modal modifiers
Lexical	Contains question words
Lexical	Contains a player's name
Lexical	Contains emoticons
Lexical	First and last words
Pattern-related	Contains a possession structure, such as *I have (no) X*
Pattern-related	Contains a query structure, such as *I need X*
Pattern-related	Contains *X for Y*
Relational	Is predicted as question wrt another speech turn
Relational	Is predicted as answer wrt another speech turn

Table 5: Feature set description

Question-answer pair		
Precision	Recall	F1 score
83.8	86.8	85.3
Q-Elab		
Precision	Recall	F1 score
53.3	57.9	55.5

Table 6: Results for the relation prediction task.

4.3 Statistical model

For our classifier, we used a regularized maximum entropy (MaxEnt, for short) model (Berger et al., 1996). In MaxEnt, the parameters of an exponential model of the following form are estimated:

$$P(b|t) = \frac{1}{Z(c)} \exp\left(\sum_{i=1}^{m} w_i f_i(t, c)\right)$$

where t represents the current dialogue turn and c the outcome (i.e., revelation of a resource or not). Each dialogue turn t is encoded as a vector of m indicator features f_i (see table 5 for more details). There is one weight/parameter w_i for each feature f_i that predicts its classification behavior. Finally, $Z(c)$ is a normalization factor over the different class labels (in this case just two, whether we have a revelation of a resource or not), which guarantees that the model outputs probabilities.

In MaxEnt, the values for the different parameters $\hat{w}$ are obtained by maximizing the log-likelihood of the training data T with respect to the model (Berger et al., 1996):

$$\hat{w} = \operatorname*{argmax}_{w} \sum_{i}^{T} \log P(c^{(i)}|t^{(i)})$$

Various algorithms have been proposed for performing parameter estimation (see (Malouf, 2002) for a comparison). Here, we used the Limited Memory Variable Metric Algorithm implemented in the MegaM package.[2] We used the default regularization prior that is used in MegaM.

4.4 Predicting the type and quantity of revealed resource

From our observations, the majority of utterances revealing resources fall into one the following two categories:

- Self-contained: resource and quantity can be deduced from the utterance alone, such as *I have no ore*;

[2] Available from `http://www.cs.utah.edu/~hal/megam/`.

Type	Keywords
Negation	no, not, don't
Second-person	you, someone, anyone
Possession	got, have, give, spare, offer
Query	want, need, get
For	for

Table 7: Markers used in type prediction

- Contextual: some information is deduced from another utterance. Both usually form a question-answer pair, such as *Do you have any wheat ? – Yes*.

We created five marker categories, described in Table 7, from the most frequent words appearing in revealing utterances. We designed a rule-based model using these markers; their combination allows us to pinpoint where the resource the player reveals is mentioned. For example, in the utterance *anyone has sheep for ore?*, the second-person marker *anyone* and the possession marker *has* indicate that the first mentioned resource is the one wanted by the player, which he doesn't reveal as possessing. Moreover, the presence of a *for* marker indicates that the players offers a resource. Hence, the resource following the marker, *ore*, is possessed by the player.

Such a rule system allows us to analyze a single utterance. However, in the case of a QAP, we often fail to retrieve data from the answer utterance alone. A second pass is thus performed on the question utterance, giving us enough context to deduce revealed resources. For example, in the QAP *anyone have wood ? – none, sorry*, in the second utterance, the negation marker *none* implies the absence of an unknown resource. The processing of the first utterance reveals that *wood* is requested by another player. We conclude that the answering players possess no wood.

We first tested our rule model on reference data, knowing exactly (from the annotations) which speech turns contain revealed resources, and which discourse relations link them. We then used the model on predicted data (discourse relations as well as dialogue turns representing revealed resources), effectively creating a full end-to-end system.

Baseline (accuracy : 82.1)			
	Precision	Recall	F1 score
H_+	54.7	73.7	.628
H_-	92.5	84.2	.882
Our method (accuracy : 89.2)			
	Precision	Recall	F1 score
H_+	75.2	70.6	.728
H_-	95.2	94.0	.933

Table 8: Results for the task of deterring whether a turn reveals a resource. H_+ represents the hypothesis that the dialogue turn does reveal a resource, while H_- the hypothesis that it doesn't.

4.5 Results

The classifier was trained using 10-fold cross-validation. For every training round, we partition the data by dialogues. With the speech turns belonging to 90% of them, we form the model training set. The turns from the remaining 10% are used as test data. We compared our method to a baseline, which does not involve machine learning. This naive model predicts revealed resource whenever a resource is mentioned by name in the utterance.

After performing ten rounds of cross-validation on the training data, we achieve a F1 score of **0.72** for the positive hypothesis *"This speech turn reveals a resource"*. The opposite class ("There is no revealed resource in this turn") has an F1 score of 0.93, achieving thus a global accuracy of 89.2%. Detailed results for our model and baseline are shown in Table 8.

Results for the prediction of resource type quantity interval are shown in table 9. As we can see, prediction of the type of resource that a player's dialogue turn reveals has an accuracy of 77% on the manually annotated instances, which falls down to 61.5% when using the results of the first classifier as input. Interval prediction on the other hand has has an accuracy of 79.9% when using manually annotated results which falls down to 65.7% when using the results of the first classifier as input. Note as well that we have implemented a baseline for both systems. Concerning resource type, the baseline randomly attributes a resource to utterances labeled as revealing one. The baseline for interval prediction assigns the most frequent interval. Results are also shown in table 9.

In table 10 we report results on the pipeline combining the three tasks. The accuracy of 57.1% does not include the instances that have been classified as not revealing any resources by the first classifier. When we evaluate both classes the accuracy goes up to 86.3%.

Accuracy	on manual annotations	on the output of the first classifier
	Baseline	
Resource type	0.165	0.146
Interval	0.559	0.328
	Our method	
Resource type	0.770	0.615
Interval	0.799	0.657

Table 9: Baseline and evaluation of predicting resource type and interval.

	Accuracy
On all instances	0.863
Only on instances classified as revealing a resource	0.571

Table 10: Results of the pipeline, that is prediction of the exact triplets **`(resource, [lower bound, upper bound])`**.

4.6 Discussion

The first step of our prediction process, locating turns revealing resources, yields very encouraging results (see Table 8): we are able to retrieve such turns with an F1 score of over 0.72, while they represent only 21% of all speech turns. On the other hand our system does not perform very well on the detection of resource type as well as the associated interval. This is to be expected: since we have split our system in three parts, there is error propagation in the pipeline. On the other hand jointly predicting the triplets is not a viable solution either, since this would lead to a great number of classes (six as we have mentioned above, multiplied by all the possible values for lower and upper bounds). We would like though to note that we greatly outperform both baselines for each of the last two tasks.

One way to improve the quality of our prediction would be to add more relational features. As context plays a critical part in determining the meaning of an utterance, features associated to its relational neighbors should be taken into account. This is true for the prediction of whether a dialogue turn reveals a resource as well as for the prediction of its type.

Accuracy for this last task is not very satisfying. The main reasons for this, which can serve as the basis for future improvements, include:

- Ambiguous *for* patterns. The utterance ***`X`*** *for* ***`Y`*** can be interpreted two ways : either as a revealing possession of **`X`** or **`Y`**. This is ambiguous even for the players themselves since often they pose a clarification question. Observation shows that the latter (possession of **`Y`**) is more frequent. The rule model implements this behavior as default when encountering such a pattern. In actual dialogues, this ambiguity is resolved by a follow-up question (*Which one are you offering ?*) or by the game context (dice rolls and resource distribution) which we haven't access to.

- Long-distance resource anaphora. On most trade negotiations, the resource being traded isn't mentioned by name at every point of the discussion, but rather referred to implicitly. When this carries over several speech turns, it becomes increasingly difficult to determine the traded resource (solving the anaphora) from a later utterance. Incorporating anaphora resolution could definitively improve our results.

- Uncommon idioms. Some utterances, such as *I'm oreless*, or *I just discarded all of my sheep*, employ rare vocabulary (with respect to the corpus) to describe resource possession. Incorporating more lexical information is necessary.

5 Related Work

Work on dialogue has traditionally focused on spoken dialogue and especially on the modeling of spoken dialogue acts (Stolcke et al., 2000; Bangalore et al., 2006; Fernández et al., 2005; Keizer et al., 2002). Recently a growing interest has emerged in working with written dialogues which can take the form either of a synchronous communication (two or multiparty live chats) or asynchronous communication (fora, email exchanges, etc). (Joty et al., 2013) are focused on the detection and labeling of topics within asynchronous

discussions, more specifically email exchanges and blogs, using unsupervised methods. (Tavafi et al., 2013) are focused on the supervised learning of dialogue acts in a broad range of domains including both synchronous and asynchronous communication. They use a multi-class SVM approach as well as two structured prediction approaches (SVM-HMM and CRFs). (Wu et al., 2002) are interested in the prediction of dialogue acts in a multi-party setting. (Joty et al., 2011) focus on the modeling of dialogue acts in asynchronous discussions (emails and fora) using unsupervised approaches. Finally, (Kim et al., 2012) are interested in the classification of dialogue acts in multi-party live chats, using a naive Bayes classifier.

Revealing a resource can be viewed as a commitment by a player that she possesses a specific resource. Public commitments have been extensively studied from a theoretical point of view in linguistics (Asher and Lascarides, 2008a; Asher and Lascarides, 2008b; Lascarides and Asher, 2009) or elsewhere (Prakken, 2005; Bentahar et al., 2005; Chaib-draa et al., 2006; Prakken, 2006; El-Menshawy et al., 2010). As far as automatic detection of public commitments is concerned, in either synchronous or asynchronous conversation, to the best of our knowledge this is the first work to explore this issue. The closest work to our own is that of (Cadilhac et al., 2013) who use the live chats from the game of *The Settlers of Catan* as well. It is concerned with the detection of dialogue acts, the detection of the resources that are givable and receivable, as well as the predictions of players' strategic actions via the use of CP-nets.

6 Conclusions

Developing a strategy in any kind of game requires reasoning about the opponents' strategies. In a win-lose game, such as the board game *Settlers* on which our experiments were based, a crucial ingredient in reasoning about everyone's strategies, including one's own, is beliefs about what resources each player possesses. Information about their resources can be inferred from observable non-verbal actions, such as the dice roll that starts each turn. Here, we provided a model for inferring information about possessed resources from verbal actions in a non-cooperative setting, where players have an incentive to conceal such information.

Our model divided the task into a three subtasks: (a) first, identify whether a dialogue turn reveals the speaker to possess a specific resource, or not; and, if so (b) identify the type of that resource, and (c) its quantity. We addressed task (a) using a statistical model of logistic regression, achieving overall accuracy of 89.2% with an F-score for the positive class (the turn reveals possession information) of 72.8% (in spite of this class comprising only 21% of the data). Our prediction of the type resource possessed and its quantity was achieved through a symbolic model, since the number of classes (5 resources, unlimited quantity intervals) makes training on the available data too sparse. While there is clearly room for improvement (61.5% accuracy on resource type; 65.7% accuracy on their quantity), our models beat a random baseline for estimating the resource type and the frequency baseline for predicting its quantity.

As we mentioned earlier, game simulations using an existing *Settlers* agent from Guhe and Lascarides (2014) show that the agent benefits if all the players' resources are made observable to him. But that's not the realistic scenario for this game, and the *Settlers* agent from (Guhe and Lascarides, 2014) for whom resources aren't made observable doesn't use the negotiation dialogues as any evidence at all about possessed resources. Instead, the agent relies only on dice rolls, build actions and robbing to update his beliefs, and so by ignoring conversation the agent can miss crucial evidence for who has what. In future work, we plan to enhance the belief model of the *Settlers* agent from Guhe and Lascarides (2014) to exploit our (noisy) model for mapping conversation to the players' resources, and evaluate whether this richer source of evidence for inferring the hidden aspects of the game state improves the agent's performance, both for successfully negotiating and winning the overall game.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cédric Degremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Vladimir Popescu, Verena Rieser, and Laure Vieu. 2012. Modelling strategic conversation: model, annotation design and corpus. In *Workshop on the Semantics and Pragmatics of Dialogue, Paris, France*. Université Paris 7, septembre.

Nicholas Asher and Alex Lascarides. 2003. *Logics of*

Conversation. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

Nicholas Asher and Alex Lascarides. 2008a. Commitments, beliefs and intentions in dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial)*, pages 35–42, London.

Nicholas Asher and Alex Lascarides. 2008b. Making the right commitments in dialogue. In *Workshop on Implicatures*. University of Michigan Linguistics and Philosophy.

Srinivas Bangalore, Giuseppe Di Fabbrizio, and Amanda Stent. 2006. Learning the structure of task-driven human-human dialogs. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 201–208. Association for Computational Linguistics.

Jamal Bentahar, Bernard Moulin, and Brahim Chaib-draa. 2005. Specifying and implementing a persuasion dialogue game using commitments and arguments. In Iyad Rahwan, Pavlos Moratis, and Chris Reed, editors, *Argumentation in Multi-Agent Systems*, volume 3366 of *Lecture Notes in Computer Science*, pages 130–148. Springer Berlin Heidelberg.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 357–368, Seattle, Washington, USA, October. Association for Computational Linguistics.

Brahim Chaib-draa, Marc-Andr Labrie, Mathieu Bergeron, and Philippe Pasquier. 2006. Diagal: An agent communication language based on dialogue games and sustained by social commitments. *Autonomous Agents and Multi-Agent Systems*, 13(1):61–95.

Mohamed El-Menshawy, Jamal Bentahar, and Rachida Dssouli. 2010. Modeling and verifying business interactions via commitments and dialogue actions. In Piotr Jdrzejowicz, NgocThanh Nguyen, RobertJ. Howlet, and LakhmiC. Jain, editors, *Agent and Multi-Agent Systems: Technologies and Applications*, volume 6071 of *Lecture Notes in Computer Science*, pages 11–21. Springer Berlin Heidelberg.

Raquel Fernández, Jonathan Ginzburg, , and Shalom Lappin. 2005. Using machine learning for non-sentential utterance classification. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 77–86.

Markus Guhe and Alex Lascarides. 2014. Game strategies in the settlers of catan. In *Proceedings of the IEEE Conference in Computational Intelligence in Games (CIG)*, Dortmund.

Shafiq R. Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1807–1813.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2013. Topic segmentation and labeling in asynchronous conversations. *Journal of AI Research (JAIR)*, 47:521–573.

Simon Keizer, Rieks op den Akker, and Anton Nijholt. 2002. Dialogue act recognition with bayesian networks for dutch dialogues. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*, pages 88–94, Philadelphia, July. Association for Computational Linguistics.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2012. Classifying dialogue acts in multi-party live chats. In *26th Pacific Asia Conference on Language,Information and Computation*, pages 463–472.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26(2):109–158.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Workshop on Natural Language Learning*, pages 49–55, Taipei, Taiwan.

M. J. Osborne and A. Rubinstein. 1994. *A Course in Game Theory*. MIT Press.

Henry Prakken. 2005. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15(6):1009–1040.

Henry Prakken. 2006. Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, 21:163–188, 6.

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.

Maryam Tavafi, Yashar Mehdad, Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2013. Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 117–121, Metz, France, August. Association for Computational Linguistics.

K. Teuber. 1995. *Die Siedler von Catan: Regelheft.* Kosmos Verlag, Stuttgart, Germany.

Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler, and William M. Pottenger. 2002. Posting act tagging using transformation-based learning. In *Foundations of Data Mining and knowledge Discovery*, pages 319–331. Springer.

Yoam Sholam and Kevin Leyton-Brown. 2009. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

Indirect answers as potential solutions to decision problems

Jon Scott Stevens & Anton Benz
Center for General Linguistics
Berlin, Germany

Sebastian Reuße, Ronja Laarmann-Quante & Ralf Klabunde
Ruhr University Bochum
Bochum, Germany

Abstract

We present a game-theoretic model of an exchange between a sales agent—an expert with access to a database of information—and a customer who poses yes/no questions to the sales agent in order to help resolve a decision problem. We first provide a game-theoretic description of such an exchange, whereby the sales agent selects an answer to the customer's question by reasoning about a space of plausible underlying decision problems. We propose a model of both answer generation and interpretation which specifies a solution to this game. The model appropriately selects indirect answers and implicatures for a particular class of yes/no questions. Implicatures can be drawn even when the speaker and hearer have partially misaligned preferences, as long as there is no incentive to lie.

1 Introduction

Indirect answers to yes/no questions come in different flavors: they may entail the direct answer, either semantically as in answer (a) in (1) or when combined with contextually shared world knowledge as in answer (b), or they may allow a direct answer to be inferred probabilistically (de Marneffe et al., 2009) as in answer (c).

(1) Q: Does the apartment have a garden?
A: a. Apartments in this neighborhood never have gardens.
b. There's no direct sunlight.
c. Gardens are pretty rare here.

Perhaps more surprising are felicitous answers whose denotation does not entail (semantically, contextually or probabilistically) a direct answer. An example of this is given in (2) in the form of an exchange between a customer looking to rent an apartment and a real estate agent tasked with helping her find the right one.

(2) CUSTOMER:
Does the apartment have a garden available?
REAL ESTATE AGENT:
It has a beautiful balcony.

Although there is nothing about the semantics of the real estate agent's response that directly suggests whether there is a garden, the real estate agent's answer is felicitous under the shared assumption that customers who are interested in a garden might also have their needs met by a balcony. For instance, the customer may want an apartment with a place to grow flowers, in which case a balcony could substitute. The real estate agent's answer implicates that the answer to the customer's question is 'no', but that the attribute supplied (that there is a balcony) serves as a substitute.

Indirect answers (and indirect speech acts in general, see e.g. Briggs and Scheutz, 2013) can reflect the answerer's ability to make inferences about the questioner's *plan* (Allen and Perrault, 1980; Green and Carberry, 1999), that is, how the current question fits into the questioner's method of accomplishing some goal. In (2), the real estate agent may guess that that goal is to find an apartment with a place to grow flowers, in which case a 'no' answer might prompt the follow-up, "does it have a balcony, then?" By looking ahead into this plan, the real estate agent can more efficiently help the customer accomplish her goal.

Also, Benz et al. (2011) suggest that recommender systems (such as a sales agent recommending objects to a customer) can exploit conceptual similarity, i.e. that such a system does better to suggest semantically related alternatives than simply to admit that the customer's needs cannot be perfectly met. For example, a customer asking for a red sofa might be recommended an orange sofa instead of simply being told that there are no red sofas on hand, or being recommended a white one.

Neither relevance to a plan nor conceptual similarity can be the whole story, however. Firstly, in a dialogue like the one in (2), the sales agent doesn't know the underlying reason for the customer's question (i.e. the customer's plan), and in fact, a number of potential reasons are plausible. Perhaps the customer wants a place to relax in the sun, or perhaps the customer specifically requires a garden. The real estate agent in this case must be able to reason probabilistically about the space of possible intents, such that (a) and (b) are possible answers in (3), but crucially not (c).[1]

[1]The awkwardness of bringing the customer's attention to the basement is mitigated in cases where a direct an-

(3) Q: Does the apartment have a garden?
A: a. It has a beautiful balcony.
b. There is a park very close by.
c. #It has a basement with a large storage area.

Secondly, semantic/conceptual similarity is not a sufficient constraint on indirect answers of this type. While such a constraint could indeed rule out the "basement" answer in (3)—'garden' and 'balcony' have many properties in common which are not shared with 'basement'—there must be more to the story, as the following example shows.

(4) Q: Is there an elementary school nearby?
A: #There is a university nearby.

Both 'elementary school' and 'university' are educational institutions, arguably as semantically related as 'garden' and 'balcony', at least in terms of their basic attributes. A better generalization is that the answer in (4) is inappropriate because elementary schools and universities do not overlap with respect to the problems they solve. In other words, the likelihood that a close-by elementary school and a close-by university will both equally satisfy the customer is simply too low for 'university' to be considered as a substitute. They don't solve the same problem for the customer.

In this paper we provide a formal model of the generation and interpretation of indirect answers of the type seen in (3). Under this model, answers are generated by reasoning about plausible motivations for asking the current question by representing a space of *decision problems* for the questioner (van Rooij, 2003; Benz and van Rooij, 2007). A speaker S (the real estate agent in our example) provides an answer to a question q, which was posed previously by a hearer H (the customer), and based on the answer to q, H chooses a resolution to her decision problem d. This process is modeled as a variant of a *signaling game* (Lewis, 1969) which generates an answer to a question together with a pragmatic interpretation and subsequent decision on the part of the hearer.

We do not assume that the goals of the speaker and the hearer are perfectly aligned. In fact, we assume that the speaker wants to steer the hearer toward a particular action (in this case continuing to consider what the salesperson is offering), and that the speaker is only cooperative in the sense that she has no incentive to tell an outright lie. By deriving the correct interpretations for the indirect answers in (3) from such a model, we show that it is possible for implicatures to arise in non-cooperative situations (see e.g. Asher and Lascarides, 2013, for a discussion of such situations), as long as honesty is enforced, either by reputation or other factors.

The remainder of this section introduces the notion of *decision problem* used in our analysis. Section 2 develops a game-theoretic description of a sales dialogue exchange, a solution to which can be calculated via an answer generation model which is given in Section 3 and an implicature calculation model which is given in Section 4. Section 5 derives the facts seen in (3) using the current model, and Section 6 concludes with a discussion of possibilities for further research.

Decision problems A decision problem is taken to be a tuple $\langle \Omega, A, U \rangle$, where Ω is a set of possible worlds (where the identity of the real world is unknown to the decider, which for our purposes is the customer), A is the set of possible actions from among which the agent must decide, and U is a utility function encoding the payoff for choosing a particular action in A given a world in Ω. The deciding agent must make inferences about the identity of the real world in order to choose the action from A which is the best candidate to maximize payoff. For current purposes we limit the space of decision problems to those that are in the real estate domain, as in (3), where there is a current "apartment under discussion", whose attributes are represented in a database visible only to the real estate agent, and where a unit of dialogue consists of a question-answer sequence pertaining to an attribute of the current apartment under discussion.[2] We assume the following correspondences for this domain.

- A "world" ω corresponds to an apartment, represented as a matrix of attribute values (e.g. $+$balcony, $-$garden, etc.) that describe the apartment.
- A consists only of two possible actions: CONTINUE and REJECT, where to CONTINUE is to carry on discussing ω and where to REJECT is to ask to end the discussion of ω.
- U assigns a utility of 1 to CONTINUE and 0 to REJECT in worlds in which CONTINUE is preferred, and assigns 0 to CONTINUE and 1 to REJECT in worlds where REJECT is preferred.

For our purposes, all possible decision problems share these constraints, such that two decision problems d

swer is also supplied, especially with a particular intonation and some hedging, e.g. "no, there's no garden, unfortunately... but there is a huge basement with lots of storage!" Although the offering of the basement attribute is directly related to the "no" answer to the customer's question in this case, it can be seen as a separate speech act and not itself the answer to the question. We are interested here only in the case where (c) is taken as an indirect answer.

[2]This simplified dialogue structure is based on observations from a series of simulated sales dialogues which we conducted between research assistants (trained to use database software and told to play the part of a sales agent) and undergraduate subjects who were instructed to find an apartment in Berlin for a hypothetical friend given some presupplied preferences.

These dialogues provided the inspiration for the data in (1)–(4); we observed that our "sales agent" readily used indirect strategies like those seen in (a) and (b) in (3), and that our "customers" had no problem interpreting them.

ω	$U(\cdot, \text{CONT.})$	$U(\cdot, \text{REJECT})$
[+garden, +balcony]	1	0
[+garden, −balcony]	1	0
[−garden, +balcony]	1	0
[−garden, −balcony]	0	1

Table 1: A decision problem d = 'ω has a place to grow flowers'

and d' differ only in which subset of the space of possible apartments determines Ω and in the binary values assigned by the utility function U. A possible utility function for a decision problem is represented in Table 1. In plain English, the decision problem represented in Table 1 corresponds to the decision on the part of H between continuing to discuss vs. rejecting the current apartment under discussion (ω) given the requirement that H must have a place to grow flowers in her new apartment.

An easier way to represent such a decision problem is as the set of worlds in which $U(\cdot, \text{CONTINUE}) = 1$. For the problem in Table 1 this is the set {[+garden, +balcony], [+garden, −balcony], [−garden, +balcony]}. Taking propositions to be sets of worlds, this is equivalent to the proposition, 'ω has a garden *or* ω has a balcony', which is in turn equivalent to the proposition (under some contextual restrictions), 'ω has a place to grow flowers.'

As mentioned above, the sales agent in (3) does not have direct access to the customer's decision problem d, and thus must reason about likely candidates for d when evaluating the felicity of an indirect answer. Therefore the sales agent must represent a space of plausible decision problems. This provides a way of encoding the sales agent's prior world knowledge—she must know in advance that the customer may want to grow flowers, relax outside, etc. Therefore, where the attributes +garden and +balcony are strongly related in virtue of belonging to at least two plausible decision problems ('ω has a place to grow flowers' and 'ω has a place to relax outside'), there is no such relation between +garden and +basement insofar as the agent cannot imagine a plausible underlying decision problem corresponding to 'ω has a garden *or* ω has a basement.' This should rule out the answer "it has a basement with a large storage area" as a possible indirect answer in (3).

2 Dialogue game

The possible space of indirect answers in a sales dialogue exchange like in (3) can be derived by first representing such an exchange as a signaling game $\mathcal{G}$, one branch of which is represented in Fig.1, equal to the tuple $\langle \{S, H\}, \Omega, \mathcal{D}, \Delta, \mathcal{Q}, M, [\![\cdot]\!], A, U_S, \mathcal{C}, U_H \rangle$ where:

- S and H are the speaker (i.e. sales agent) and hearer (customer), respectively.
- Ω is the set of possible worlds, where a world is conceived of as an attribute value matrix exhaustively specifying the attributes of a single possible database object.
- $\mathcal{D}$ is the set of shared plausible decision problems, each represented as the set of worlds (i.e. proposition) in which the best decision for the hearer is to continue discussing the current database object ω. ($\mathcal{D} \subset \mathcal{P}(\Omega)$.)
- Δ, a function from $\Omega \times \mathcal{D}$ to the interval $[0, 1]$, is a probability distribution over worlds and decision problems. We assume that Δ is flat, i.e. worlds and decision problems are *a priori* equiprobable, and that Δ provides prior probability terms for Bayesian posterior probabilities which determine expected utility for the speaker and hearer.
- $\mathcal{Q}$ is the set of possible attribute queries, e.g. questions of the form 'what is the value of attribute α in ω?', where each question is conceived of as a set of possible answers (Hamblin, 1973), or a set of sets of worlds (a set of worlds being a proposition). ($\mathcal{Q} \subset \mathcal{P}(\mathcal{P}(\Omega))$.)
- M is a language of possible *messages*, in this case taken to be the set of possible answers to an attribute query.
- $[\![\cdot]\!]$ is a denotation function, from M to $\mathcal{P}(\Omega)$.
- A is the set of possible hearer actions, equal to the set {CONTINUE, REJECT}.
- U_S is a function from $\Omega \times M \times A$ to the interval $[0, 1]$, specifying speaker utility.
- $\mathcal{C}$ is a function from M to the interval $[0, 1]$ corresponding to the *cost* of sending a message in M. We assume a higher cost for longer messages and a nominal cost for not providing (the semantic equivalent of) a literal yes/no answer to the hearer's question q (i.e. a member of q).

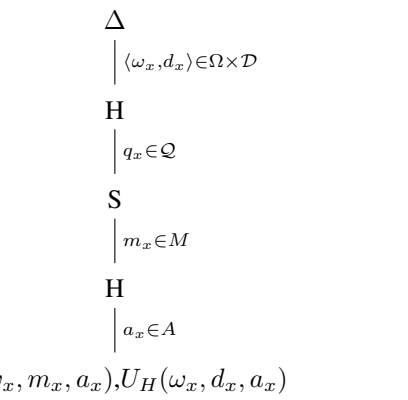


Figure 1: A branch of $\mathcal{G}$ (ω_x is unknown to H, and d_x is unknown to S)

- U_H is a function from $\Omega \times \mathcal{D} \times A$ to $\{0, 1\}$, specifying hearer utility.

U_S and U_H represent imperfectly aligned preferences on the part of the speaker and hearer, such that: (i) the hearer's utility is positive only if she continues discussing an apartment which solves her decision problem or rejects one which doesn't, and (ii) the speaker's utility is positive only if the hearer chooses to continue. This reflects the fact that in many sales dialogues, the sales agent has strong incentive to sell a particular object, e.g. if it is expensive and she works on commission. Honesty is strictly enforced, encoding a strong role for reputation in the possible answers given by the sales agent. (After all, outright lying to your customers tends to be a bad business decision.) Thus, the speaker's utility function is positive only if her utterance is true. The cost term $\mathcal{C}(m)$, taken to encode both a higher cost for increased message length and a nominal cost for non-literal answers, is subtracted from base values of 1 and 0. The utility functions for the hearer and speaker, respectively, are as follows.

$$\begin{aligned} U_H(\omega, d, a) &= 1 \text{ if } \omega \in d \;\&\; a = \textsc{continue} \\ &= 1 \text{ if } \omega \notin d \;\&\; a = \textsc{reject} \\ &= 0 \text{ otherwise} \end{aligned} \tag{1}$$

$$\begin{aligned} U_S(\omega, m, a) &= 1 - \mathcal{C}(m) \text{ if } \omega \in [\![m]\!] \;\&\; a = \textsc{cont.} \\ &= -\mathcal{C}(m) \text{ otherwise} \end{aligned} \tag{2}$$

U_S does not depend on the hearer's decision problem d directly, but it depends on the hearer's action, which is in turn dependent on d. Therefore, S will indeed need to reason probabilistically about d (which is unknown to S) in order to choose an optimal message. The probability of any particular d can be inferred on the basis of the question q. Rather than supplying the model an externally determined probability distribution over decision problems, we assume that all decision problems are *a priori* equiprobable, and that S can infer a conditional probability function over decision problems, $P(\cdot|q)$, via Bayesian reasoning. Bayes' theorem specifies $P(d|q)$ as the product of $P(q|d)$ and the fraction $P(d)/P(q)$. Assuming $P(d)$ and $P(q)$ to be constants, $P(d)/P(q)$ serves only to normalize the values of $P(q|d)$ for all d in $\mathcal{D}$. We take $P(q|d)$ to be the probability of randomly selecting q from the set of simple attribute queries (i.e. the subset of $\mathcal{Q}$) which has the property that at least one answer in that set, if true, would solve d (i.e. make one action dominant over the other). This can be formulated as follows, where the numerator returns 1 iff q contains an answer which solves d and 0 otherwise (the "int" function transforms boolean values into 0 or 1), and where the denominator is the size of the set of all questions in $\mathcal{Q}$ that contain such an answer.

$$P(q|d) = \frac{\text{int}(\exists \phi \in q.\ \phi \subseteq d)}{|\{q \subset \mathcal{Q} | \exists \phi \in q.\ \phi \subseteq d\}|} \tag{3}$$

In addition to representing a probability for each possible d, the speaker must also have a belief for each possible d about which action a type-d hearer will take given the content of the speaker's message—this is the hearer's *strategy* for selecting an action. By first assuming some fixed strategy for the hearer, the speaker can determine which message has the best chance of leading to an outcome which maximizes the speaker's own utility. Since the hearer may choose an action at random in some situations, the hearer's strategy is represented as a probability distribution over actions, $\mathcal{H}(a|d, m)$. We can now specify an expected utility function for the speaker, which returns the weighted average, for all possible underlying decision problems, of the expected payout to the speaker given that decision problem and the hearer strategy $\mathcal{H}$.

$$EU_S(\omega, m|q, \mathcal{H}) = \sum_{d \in \mathcal{D}} P(d|q) \cdot \sum_{a \in A} \mathcal{H}(a|d, m) \cdot U_S(\omega, m, a) \tag{4}$$

Similarly, expected utility for the hearer is calculated by assuming a fixed strategy for the speaker. The posterior probability $P(\omega|m, \mathcal{S})$ assigns zero probability to any world in which the speaker would not send m assuming some fixed speaker strategy $\mathcal{S}$, where $\mathcal{S}(\omega, q)$ outputs a message.

$$EU_H(d, a|m, \mathcal{S}) = \sum_{\omega \in \Omega} P(\omega|m, \mathcal{S}) \cdot U_H(\omega, d, a) \tag{5}$$

The optimal behavior in a dialogue exchange like the one in (3) is specified by an equilibrium in $\mathcal{G}$, which is a pair of strategies $\langle \mathcal{S}, \mathcal{H} \rangle$ such that each player's expected utility is maximized by playing their own strategy while assuming the other player's strategy to be fixed. (In other words, no single player does better by unilaterally deviating from $\langle \mathcal{S}, \mathcal{H} \rangle$.)

We now propose an answer generation procedure for the speaker (sales agent) which specifies a strategy $\mathcal{S}$ which is part of an equilibrium in this game.[3] This generation model is shown to correctly predict constraints on indirect answers for a fragment of sales dialogue.

3 Indirect answer generation

Given the game $\mathcal{G}$ introduced in the previous section, an optimal answer for the sales agent in a dialogue exchange of this type is one that maximizes the odds that the customer will be prompted to choose the action CONTINUE. Given the utility structure for $\mathcal{G}$, a rational customer will choose CONTINUE if the denotation of the hearer's message is a subset of d. (A rational customer assumes the message to be true, knowing there is no incentive to lie in this situation.) If the

[3] We include no formal proof here due to space constraints, but it can be shown that the speaker and hearer strategies given in the following two sections correspond to a perfect Bayesian equilibrium (Harsanyi, 1968; Fudenberg and Tirole, 1991).

customer's underlying decision problem d were known, the speaker's problem would reduce to that of finding the least costly true message for which this holds. Of course d is not known, and so probabilistic reasoning must be incorporated into the speaker's strategy. To this end, we first define a set $\mathcal{D}'_m$ of "compatible decision problems" given a message m.

$$\mathcal{D}'_m = \{d \in \mathcal{D} \mid [\![m]\!] \subseteq d\} \tag{6}$$

The speaker does best by maximizing the *probability of compatibility* (P_{comp}) between a given message m and whichever value of d holds for the hearer.

$$P_{comp}(q, m) = \sum_{d \in \mathcal{D}'_m} P(d|q) \tag{7}$$

The optimal answer for the speaker, then, is a true message which maximizes P_{comp} and minimizes cost.

We assume that the cost function $\mathcal{C}(m)$ grows with the size of the message such that the speaker prefers messages which convey a single attribute of the database object under discussion. Without such an assumption, the optimal message would always be to list all possible solutions to the hearer's underlying decision problem, rather than choosing one alternative over another, a strategy which seems to be rare in real dialogue situations. Although relatively short conjunctive answers to (3) such as "it has a beautiful balcony, and there is a park nearby" are not infelicitous, we consider for simplicity's sake only a set $M' \subset M$ of messages which convey a single attribute.

Also, recall that $\mathcal{C}(m)$ encodes a nominal (i.e. tie-breaking) cost for indirect answers such that, if all options are otherwise equal, the speaker prefers simply to provide a literal yes/no answer. This cashes out the intuition that, if the speaker is guaranteed to lose utility by responding to q, that is, if the object under discussion has no chance of being desirable to the customer given her decision problem, the speaker wishes to appear cooperative by providing a direct answer, e.g. "unfortunately there is no garden" over an irrelevant response, e.g. "my sister paints portraits of bees", which would otherwise yield the same utility for the speaker. Like the enforcement of honesty, this could be seen as a byproduct of reputation, or instead seen as a reflex of coherence requirements or discourse obligations which are introduced by the question (Traum and Allen, 1994).

Putting it all together, the optimal speaker strategy in the game $\mathcal{G}$ is obtained via the following answer generation procedure, for which we first give an informal specification.

1. Let M'_{TRUE} be the subset of M' which excludes all false messages.
2. Obtain the set of messages in M'_{TRUE} that maximize the probability P_{comp} that m is compatible with (i.e. is a subset of) the hearer's underlying decision problem.
3. Eliminate from that set any messages for which there is a lower cost alternative, where a message has lower cost iff it directly answers the hearer's question q (i.e. if $[\![m]\!] \in q$).
4. Output a random message from that new set.

Formally, this can be represented with the following algorithm for a speaker strategy $\mathcal{S}(\omega, q)$, which outputs a message.

1. Let $M'_{\text{TRUE}} = \{m \in M' \mid \omega \in [\![m]\!]\}$
2. Let $\mu = \texttt{arg max}_{m \in M'_{\text{TRUE}}} P_{comp}(q, m)$
3. Let $\mu' = \{m \in \mu \mid \exists m' \in \mu.\ [\![m']\!] \in q \rightarrow [\![m]\!] \in q\}$
4. Output some member of μ'

4 Implicature calculation

One *prima facie* peculiarity with the speaker's strategy $\mathcal{S}(\omega, q)$ is that it filters potential messages by maximizing the likelihood that they will guarantee a CONTINUE action, and does not consider the possibility that a message will make the hearer indifferent between CONTINUE and REJECT, which could in some instances benefit the speaker. For example, one might argue that the answer "there is a basement" for (3) is better for the speaker than the direct answer "there is no garden" under $\mathcal{G}$, because the former is guaranteed not to address d at all, and thus would make the hearer indifferent, leading to a 0.5 probability of the desired CONTINUE outcome, whereas the latter could result in a guaranteed REJECT outcome if the hearer's decision problem is simply 'ω has a garden.' In other words, one might argue that a non-sequitur answer is better than one that might prompt a negative reaction from the hearer, even considering any nominal costs for not directly answering the current question.

This proves not to be a problem, however, because a truly rational speaker will take into account the implicatures that the hearer will draw from her message. For example, if the speaker answers the question, "does the apartment have a garden?" with "it has a basement", the hearer knows that the speaker would have been better off saying "yes" to her question if she could have done so truthfully. Therefore, that answer must be false in the current world. This implicature (that the apartment in fact does not have a garden) makes the "basement" answer equivalent to a "no" answer, except that it bears an increased cost for being a non-literal answer, i.e. for failing to provide a direct 'yes' or 'no' answer.

This is encoded in the hearer's expected utility function for $\mathcal{G}$ via $P(\cdot|m, \mathcal{S})$: if the hearer's beliefs are reasonable, then she will assign zero probability to worlds in which m is not a possible output of $\mathcal{S}(\omega, q)$, thereby drawing the implicature that any messages that would otherwise be better for the speaker are false in ω. This should be made part of the hearer strategy $\mathcal{H}$

which specifies the space of hearer-optimal responses to m, which in turn determines EU_S, and with it the speaker's optimal message. Because the speaker considers $\mathcal{H}$, the speaker knows that an alternative speaker strategy $\mathcal{S}'$ which attempts to trick the hearer with non-sequiturs, is necessarily less optimal than $\mathcal{S}$.

The aforementioned implicatures[4], which can serve to provide a direct answer to the hearer's question, can be calculated by reverse engineering the speaker's strategy and assuming the falsity of messages that would be more optimal than the observed one if true. This can be accomplished by simply assuming the falsity of any message which has a higher value for P_{comp} than the message that was actually sent. This yields the following algorithm, which we'll call IMPL, which takes a message m as input and outputs a proposition.

1. Let $\beta = \{[\![m']\!] \in M' \mid P_{comp}(q, m') > P_{comp}(q, m)\}$
2. Output $\Omega \setminus \cup\beta$

This outputs only the implicatures drawn from m; the complete pragmatic interpretation assigned to m by the hearer is $[\![m]\!] \cap \text{IMPL}(m)$. The hearer's strategy, then, can be specified as follows.

$$\begin{aligned}\mathcal{H}(\text{CONT.}|d, m) &= 1 \text{ iff } [\![m]\!] \cap \text{IMPL}(m) \subseteq d \\ &= 0 \text{ iff } [\![m]\!] \cap \text{IMPL}(m) \cap d = \emptyset \\ &= 1/2 \text{ otherwise} \\ \mathcal{H}(\text{REJECT}|d, m) &= 1 - \mathcal{H}(\text{CONT.} \mid d, m)\end{aligned} \tag{8}$$

5 Example

We now use the answer generation and implicature calculation procedures given above to derive the facts in (3), reproduced below as (5), given a fragment of world knowledge.

(5) H: Does the apartment have a garden?
S: a. It has a beautiful balcony.
b. There is a park very close by.
c. #It has a basement with a large storage area.

Although a decision problem is formally represented as the set of worlds in which the decision problem is solved, any decision problem consistent with the sales agent's world knowledge can also be represented as a complex preference statement, e.g. 'ω has a balcony *or* ω has a garden.' While conjunctive decision problems are logically possible, we only consider disjunctive ones, i.e. decision problems that can be phrased as 'ω has value x for attribute α *or* ω has value y for attribute β.' Accordingly, we use a short-hand set notation, such that $\{+\alpha, +\beta\}$ means the proposition 'ω is $+\alpha$ *or* ω is $+\beta$.' Using this notational shortcut, we begin to build a fragment of world knowledge with which to derive example (5).

Consider a fragment of a context for example (5) where there are only four apartment attributes represented in the database: (i) whether there is a garden available, (ii) whether there is a balcony, (iii) whether there is a park nearby[5], and (iv) whether there is a basement storage area available. To abbreviate, we use 'B' for *balcony* and 'K' (as in German *Keller* 'basement') for *basement*. Table 2 shows the possible worlds.

The space of possible questions is: $\mathcal{Q} = \{$'What is the value for attribute α in ω?'$\}$, where $\alpha \in \{$garden, balcony, park, basement$\}$, and the current question under discussion is $q =$ 'What is the value for attribute *garden* in ω?', equivalent to the set containing: (i) the set of worlds in which ω has a garden, and (ii) the set of worlds in which ω does not have a garden.

Table 3 shows the decision problems deemed to be reasonable in this fragment, along with their conditional probabilities. We consider the following possibilities: the customer either wants a garden, balcony, park or basement specifically, or else a place to grow flowers, a place nearby to go for a walk outside, or just a place to relax outside.

Table 4 specifies a space of possible utterances, all specifying a +/− value for a single attribute. Table 5 shows binary truth values for whether m is in d for all m/d combinations, as well as the conditional probabilities for each d, the value of $P_{comp}(q, m)$ for each message, and whether each m is a literal answer (that is, whether the denotation of m is in q). Putting it all together, we obtain the following dominance hierarchy of best messages. The speaker should use the best message that also happens to be true.

$$m_G \succ m_B, m_P \succ m_{-G}$$

In plain English, we have obtained the following strategy for our sales agent for this particular dialogue exchange.

1. If ω has a garden say, "there is a garden."
2. Else, if ω has a balcony say, "there is a balcony", or if ω has a park nearby say, "there is a park nearby."
3. Else, say, "there is no garden."

Finally, we can use the hearer's representation of the speaker's strategy to derive the indirect meaning carried by the speaker's answer.[6]

[4] We emphasize that no claims are made about the generalizability of the current model to other kinds of implicatures, e.g. those which arise in purely cooperative dialogue situations.

[5] It is a simplification to treat this as a binary variable; in actuality, the database would contain a distance value to the nearest park, with the definition of "nearby" left to the judgment of the interlocutors.

[6] Note that the implicature algorithm in Section 4 assumes that the hearer only considers P_{comp}, and not the cost for the speaker. This allows the hearer to derive correct implica-

Ω	AVM	Ω	AVM
ω_{GBPK}	[+garden, +balcony, +park, +basement]	ω_{GBP}	[+garden, +balcony, +park, −basement]
ω_{GBK}	[+garden, +balcony, −park, +basement]	ω_{GPK}	[+garden, −balcony, +park, +basement]
ω_{BPK}	[−garden, +balcony, +park, +basement]	ω_{GB}	[+garden, +balcony, −park, −basement]
ω_{GP}	[+garden, −balcony, +park, −basement]	ω_{GK}	[+garden, −balcony, −park, +basement]
ω_{BP}	[−garden, +balcony, +park, −basement]	ω_{BK}	[−garden, +balcony, −park, +basement]
ω_{PK}	[−garden, −balcony, +park, +basement]	ω_{G}	[+garden, −balcony, −park, −basement]
ω_{B}	[−garden, +balcony, −park, −basement]	ω_{P}	[−garden, −balcony, +park, −basement]
ω_{K}	[−garden, −balcony, −park, +basement]	$\omega_{\emptyset}$	[−garden, −balcony, −park, −basement]

Table 2: Worlds

$\mathcal{D}$	Attributes	Plain English	$P(\cdot\|q)$
d_G	{+garden}	'Access to a garden'	6/14
d_B	{+balcony}	'A balcony'	0
d_P	{+park}	'A nearby park'	0
d_K	{+basement}	'A basement'	0
d_F	{+garden, +balcony}	'A place to grow flowers'	3/14
d_W	{+garden, +park}	'A place to walk outside'	3/14
d_R	{+garden, +balcony, +park}	'A place to relax outside'	2/14

Table 3: Plausible decision problems

M'	English	$[\![\cdot]\!]$
m_G	"There is a garden"	$\{\omega_{GBPK}, \omega_{GBP}, \omega_{GBK}, \omega_{GPK}, \omega_{GB}, \omega_{GP}, \omega_{GK}, \omega_{G}\}$
m_B	"There is a balcony"	$\{\omega_{GBPK}, \omega_{GBP}, \omega_{GBK}, \omega_{BPK}, \omega_{GB}, \omega_{BP}, \omega_{BK}, \omega_{B}\}$
m_P	"There is a park nearby"	$\{\omega_{GBPK}, \omega_{GBP}, \omega_{BPK}, \omega_{GPK}, \omega_{PK}, \omega_{BP}, \omega_{GP}, \omega_{P}\}$
m_K	"There is a basement area"	$\{\omega_{GBPK}, \omega_{BPK}, \omega_{GBK}, \omega_{GPK}, \omega_{PK}, \omega_{GK}, \omega_{BK}, \omega_{K}\}$
m_{-G}	"There is no garden"	$\{\omega_{BPK}, \omega_{BP}, \omega_{BK}, \omega_{PK}, \omega_{B}, \omega_{P}, \omega_{K}, \omega_{\emptyset}\}$
m_{-B}	"There is no balcony"	$\{\omega_{GPK}, \omega_{GP}, \omega_{GK}, \omega_{PK}, \omega_{G}, \omega_{P}, \omega_{K}, \omega_{\emptyset}\}$
m_{-P}	"There is no park nearby"	$\{\omega_{GBK}, \omega_{GB}, \omega_{GK}, \omega_{BK}, \omega_{G}, \omega_{B}, \omega_{K}, \omega_{\emptyset}\}$
m_{-K}	"There is no basement area"	$\{\omega_{GBP}, \omega_{GB}, \omega_{GP}, \omega_{BP}, \omega_{G}, \omega_{B}, \omega_{P}, \omega_{\emptyset}\}$

Table 4: Messages

	m_G	m_B	m_P	m_K	m_{-G}	m_{-B}	m_{-P}	m_{-K}	$P(\cdot\|q)$
$[\![\cdot]\!] \subseteq d_G$	1	0	0	0	0	0	0	0	6/14
$[\![\cdot]\!] \subseteq d_B$	0	1	0	0	0	0	0	0	0
$[\![\cdot]\!] \subseteq d_P$	0	0	1	0	0	0	0	0	0
$[\![\cdot]\!] \subseteq d_K$	0	0	0	1	0	0	0	0	0
$[\![\cdot]\!] \subseteq d_F$	1	1	0	0	0	0	0	0	3/14
$[\![\cdot]\!] \subseteq d_W$	1	0	1	0	0	0	0	0	3/14
$[\![\cdot]\!] \subseteq d_R$	1	1	1	0	0	0	0	0	2/14
P_{comp}	1	5/14	5/14	0	0	0	0	0	
$[\![m]\!] \in q$	1	0	0	0	1	0	0	0	

Table 5: Optimality of messages if true. An answer is sub-optimal if there is a true answer within a group to its left, as indicated by dashed lines. Within each grouping, a message is optimal only if either (i) it is a literal answer, or (ii) there are no literal answer alternatives that could be used.

1. If the speaker answers either "there is a balcony" or "there is a park nearby", then there is no garden.

2. If the speaker answers, "there is no garden", then there is no garden, balcony, or park nearby.

While the first implicature is clear from example (5), the second seems disputable. Is it really the case that a "no" answer implicates that there are no possible substitute solutions to the hearer's problem? One gets the intuition that this is only the case under strong common knowledge assumptions about how willing the sales agent is to query the database for multiple attributes to find alternatives. This willingness undoubtedly depends on personality traits which must be attributed to the sales agent by the customer (see Walker et al., 1997, and related work), for example laziness. If there is the possibility of a lazy sales agent, for example, the hearer will be less ready to draw the second implicature, because she cannot be certain that the sales agent has checked the database to see whether there is a balcony or a park nearby. But the first implicature is a safe bet in any case, because the customer can be sure that the sales agent has checked to see whether there is a garden, since that attribute was the target of the customer's question. This intuition could be cashed out within the current framework as an effect of uncertainty on the hearer's part about the cost function $\mathcal{C}$. The first implicature, but not the second, is calculable under any reasonable value for $\mathcal{C}(m)$. Further investigation of such effects must be left to future research.

6 Discussion

We have presented a game-theoretic description of a yes/no question-answer exchange between a sales agent and a customer in which the sales agent (speaker) must consider the customer's (hearer's) underlying decision problem which motivated her question before supplying an answer. We have proposed speaker and hearer strategies designed to find equilibria in this game. The resulting model has three key properties. First, the speaker has motivation to produce indirect answers insofar as those answers serve as potential alternative solutions to the hearer's underlying problem. Second, the hearer can infer a direct answer to her question from an indirect one, even if no entailment relationship exists between the speaker's response and a direct yes/no answer. Third, these inferences are possible even when the speaker and hearer have partially misaligned goals.

The partial misalignment of preferences in the model represents a move beyond traditional Gricean accounts of implicature into cases where the speaker has some incentive to be non-cooperative (what Asher and Lascarides, 2013, call "strategic conversation"). Under our model, implicatures arise in non-cooperative situations as long as honesty is enforced, either through reputation or through other means. In a sales dialogue like the one studied here, the sales agent wants the customer to choose the action CONTINUE regardless of whether the object being sold is truly optimal for the customer, and yet if she cannot lie, the sales agent behaves as if she is fully cooperative. The reason for this is that, if the salesperson's goals are known by the customer, then the customer will draw implicatures from any indirect answers by assuming the falsity of any answers that would have been more optimal given those goals. Misleading irrelevant answers become no better than answers which directly prompt an unwanted action from the customer—the customer is too smart to be swindled.

This work is intended as a starting point for a more general inquiry into such phenomena in dialogue. Further research is required to assess the generalizability of the current approach to different dialogue situations, as well as the validity of our assumptions regarding how world knowledge is represented in the dialogue model. For example, we currently posit that the interlocutors have access to a discrete space of plausible decision problems ($\mathcal{D}$), such that extremely unlikely question motivations (e.g. $d =$'ω has a place for my cat, who only likes balconies and basements, to take naps') are not considered. It is important to determine whether this aspect of our approach is fully justified, and, if so, how such a discrete space might be built and represented from prior experience.

Finally, future research will determine whether such considerations can be practically implemented within an automated dialogue system. Namely, while the algorithm in Section 3 can be used to select from among a finite space of possible answers to a yes/no question, the output relies crucially on the space of possible decision problems. It remains to be assessed whether a richer space could be empirically obtained, and whether such a space would yield realistic answers to a wider variety of questions in a sales dialogue.

tures even for "off-equilibrium" messages. For example, it is never optimal, due to cost, for the speaker to answer, "it has a basement", but if the speaker did so for some reason, the hearer would still reason that there is no balcony or park or garden. If the hearer considered cost as well, then she would reason that both "there is a garden" and "there is no garden" are false—a logical contradiction.

References

Allen, J. F. and Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.

Asher, N. and Lascarides, A. (2013). Strategic conversation. *Semantics and Pragmatics*, 6(2):1–62.

Benz, A., Bertomeu, N., and Strekalova, A. (2011). A decision-theoretic approach to finding optimal responses to over-constrained queries in a conceptual search space. In *SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 37–46.

Benz, A. and van Rooij, R. (2007). Optimal assertions, and what they implicate. a uniform game theoretic approach. *Topoi*, 26(1):63–78.

Briggs, G. and Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence,(forthcoming)*.

de Marneffe, M.-C., Grimm, S., and Potts, C. (2009). Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 136–143. Association for Computational Linguistics.

Fudenberg, D. and Tirole, J. (1991). Perfect Bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory*, 53(2):236–260.

Green, N. and Carberry, S. (1999). Interpreting and generating indirect answers. *Computational Linguistics*, 25(3):389–435.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics vol. 3*. Academic Press, New York.

Hamblin, C. (1973). Questions in Montague english. *Foundations of Language*, 10:41–53.

Harsanyi, J. C. (1968). Games of incomplete information played by 'Bayesian' players, part ii. *Management Science*, 14(5):320–334.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge University Press, Cambridge.

Traum, D. R. and Allen, J. F. (1994). Discourse obligations in dialogue processing. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 1–8.

van Rooij, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, 26(6):727–763.

Walker, M. A., Cahn, J. E., and Whittaker, S. J. (1997). Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the first international conference on autonomous agents*, pages 96–105. ACM.

Credibility and its Attacks

Antoine Venant[1], **Nicholas Asher**[2] **and Cedric Dégremont**[1*]
[1]IRIT, Université Paul Sabatier and [2]CNRS, IRIT

1 Introduction

Dialogues occurring in non-cooperative settings often exhibit attempts at deception, such as misdirections or lies. In many contexts, such as, e.g., trials or political debates, the objectives of a conversation's participant cannot be expressed in terms of her and her opponent's beliefs toward the content of the different dialog moves. In such contexts, dialogues moves come with semantic commitments of their own and challenges based on other players' semantic commitments.

In a political debate, an agent A might ask a question to another agent B, even though A knows the answer to the question. In such a case A is just seeking for B's commitment to an answer. If B complies and provides an answer, it can be in A's interest to further challenge this answer, even knowing it is correct. What is crucial here, are the objective semantic commitments that agents can force out of each other, rather than the subjective beliefs of these agents about whether the content of these commitments actually occurs or not.

Addressing the above requires us: 1. to have a semantic theory of commitments in dialogues, 2. to determine semantically what constitutes an attack and 3. to distinguish between attacks from a semantic perspective.

In the next section, we define credibility more precisely and attacks on it, linking these to linguistic commitments. In section 3, we give some examples of attacks on credibility, while sections 4 and 5 flesh out the analysis. Section 6 describes related work. We conclude with some directions for the future in section 7.

2 Credibility and commitments

An attack on credibility can be thought of as exposing deceitful intention. But determining intentions behind speech acts is a tricky business [14] we will not be getting into. The notions of credibility and attacks we are considering depends on overt and public linguistic commitments by speakers.

Using commitments, we now precise our notion of credibility: a dialogue agent i is not credible iff (i) it is shown for some φ that i has committed to φ that is absurd or clearly refutable (shown to be inconsistent with a prior claim of the agent or a background common assumption), and that it was plausibly in i's interest to commit to φ if φ is not attacked. An *attack* by player j on the credibility of i occurs iff j commits to the following: i has committed to φ, $\varphi \models \psi$, ψ is absurd or refutable, and it is in o's interest to commit to ψ, if ψ is not attacked. A move a by player i *makes possible* an attack on credibility iff it is discourse coherent for j to attach an attack on i's credibility to a.

Our notion of credibility differs considerably from that employed in the signaling games literature where credibility is defined in terms of beliefs, typically in equilibrium [10, 11]. Our notion of credibility is defined in terms of commitments, agent's interests and logical consequence, none of which depend on how the message affects the agents' beliefs.

To flesh out our picture of credibility and attack, we need to explain our notions of consequence and interest or preference. We have two notions of consequence: ordinary, logical consequence and defeasible consequence. We will assume that our agents are logically (though not factually) omniscient and so if i commits to φ he publicly also commits to ψ if ψ is a logical consequence of φ (notation $\varphi \models \psi$). Agents also commit to implicatures that are defeasible but what we shall term normal consequences that interlocutors would draw upon learning that i commits to φ. Finally, implicatures may be more tentative, as when i draws attention to an alternative to some-

*We thank ERC Research Grant 269427 for research support

thing to which he is explicitly committed. We'll assume that implicatures are modeled in a defeasible logic using a space of preferred models of the conversation. We also allow that some weak implicatures may exist only in some of the preferred models while stronger ones are true in all preferred models. We thus distinguish between the following three levels of commitment. -**Non-defeasible commitment by** i **to** φ: φ is a logical consequence of every possible interpretation of i's contribution. -**Implicit defeasible commitment by** i **to** φ: the "prefered" interpretations of i's contribution entail φ. -**Weak implicit defeasible commitment by** i **to** φ**:** some interpretations of i's contribution imply φ. Section 4 will provide more formal definitions.

We take preferences to be tied to a conception of rationality. In our framework, we will assume two conversational partners $0, 1$ and a third party who observes and judges but does not participate in linguistic exchanges. We will refer to this third party as *the jury*. The jury should be thought off as an abstract procedural entity with an objective look on the conversation that serves in the process of modeling rationality. The jury's mecanisms also depend on some contextual parameters which are common knowledge among the players: for instance, at court, the jury should know that an "honnest" expert witness must not share interest with the defendent lawyer. To give another example, he should know also when some facts are irrefutable and known as such. With this notion of a jury, our players prefer moves which make them look good in the eyes of the jury and make the other look bad, or at least worse. An attack by i on a player j's credibility is a way to make j look less good. Part of i's looking good is to not make mistakes, to not invite attacks on her credibility, but to make herself look good a player must provide positive reasons for the position she favors. *Mutatis mutandis* for the preferences of player j. More generally: (i) our players must play moves that make them look good; (ii) if player i is rational, she will prefer moves that make possible moves that j cannot attack; (iii) between 2 moves that make i look good but make possible attacks, she will prefer the one with the more indirect or weaker damaging context, since a more indirect damaging consequence is one that has a rebuttal move *that's not what I meant to say*.

3 Linguistic examples and intuitions

In this section we offer some linguistic examples featuring different sorts of commitments and attacks on credibility. These examples involve not only commitments to propositions expressed by assertoric clauses but also to propositions involving *rhetorical relations* that link clauses, sentences and larger units together into a coherent whole. That is, players commit to a particular content *and* to its relations with what has been said before. In so doing a player may also commit to contents proferred by his conversational partner as in [1].

Consider a case in which speaker A takes C's initial moves to be ambiguous.

(1) a. C: N. isn't coming to the meeting. It's been cancelled.
b. A: Did you mean that N. isn't coming because the meeting's cancelled or that the meeting is cancelled as a result?
c. C: As a result.

A's clarification question in (1)b presupposes that C's initial contribution was ambiguous between a result and an explanation move [7, 16]. We take this to imply at least a weak implicature for both readings, either of which a conversational participant could have exploited. This is something we want to model, and we'll see in the next example how such implicatures are exploited by an interlocutor.

Now consider the following example:

(2) a. C: N. isn't coming to the meeting. It's been cancelled.
b. A: That's not why N. isn't coming. He's sick.
c. C: I didn't say that N. wasn't coming because the meeting was cancelled. The meeting is cancelled because N. isn't coming.

This example illustrates how commitments embed. In (2)b A commits to the fact that C committed in (2)a to providing an explanation for why N isn't coming, even though (2)a is ambiguous. Only such a commitment explains why A attacks that commitment in the way that he does by giving an alternative explanation. But in fact, C takes that commitment by A to have misinterpreted him; C commits in (2)c that he committed in (2)a to of-

fering a consequence or result of N's not coming to the meeting.

Note that while A attacks a move of C's in (2), he does not attack C's credibility in our sense. But neither does (2) provide a case of misleading implicature. However, the following example from [3] does. During the Dan Quayle-Lloyd Bentsen Vice-Presidential debate of 1988, Quayle was repeatedly questioned about his experience and his qualifications to be President. Quayle's attempted to compare his experience to the young John Kennedy's (referred to below as *Jack Kennedy* to convey familiarity) in his answer.

(3) a. 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.
b. Bensten: Senator, I served with Jack Kennedy. I knew Jack Kennedy. Jack Kennedy was a friend of mine. Senator, you're no Jack Kennedy.

Implicatures play a key role in his example. Quayle argues, against the thesis that his little governmental experience would make him unsuitable for the presidency, that Kennedy before him, with as much experience as he have, was able to handle the presidency. But this answer to the question suggests an implicit comparison between the two politicians (both junior senators from a state, each with little governmental experience) and gives rise to the possibility of interpreting Quayle's move as a stronger commitment that he would likely be able to handle the presidency in the same way that John Kennedy handled his, which, if not challenged would serve Quayle's claim better. Bentsen seized upon this weak implicature of Quayle's contribution and refuted it, indirectly exposing to the audience the self-serving nature of the comparison.

Here's an attested example from [17], in which a prosecutor (P) wants Bronston (B) to say whether he had a bank account in Switzerland or not, and Bronston does not want to make such a commitment for strategic reasons. But he defeasibly commits to an answer with (4)d in an attempt to avoid further questioning [2].

(4) a. P: Do you have any bank accounts in Swiss banks, Mr. Bronston?
b. B: No, sir.
c. P: Have you ever?
d. B: The company had an account there for about six months, in Zurich.

It is interesting to consider a continuation of this in which the prosecutor would indirectly attack this response in (4)d.

(5) Prosecutor: I would like to know whether you personally ever had an account there?

If Bronston is forced on the threat of perjury to answer affirmatively, his response in (4)d now looks pretty deceiving to the Jury. The natural thought arises: Bronston was trying to deceive us into thinking that he didn't have an account. Though the prosecutor didn't proceed as in (5), had he done so he would have successfully attacked Bronston's credibility.

For our final example, consider the following excerpt from a *voir dire* examination in [12]. As background, the plaintiff lawyer (LP) has been repeatedly coming back to questions about the division of a nerve during a surgery with the objective of getting the witness (D) to characterize the surgical operation as incompetent and mishandled. Repeatedly coming back to the topic wore D down, and the defense attorney (LD) was no help:

(6) a. LP: And we know in addition to that, that Dr. Tzeng tore apart this medial antebrachial cutaneous nerve?
b. D: Correct.
c. LD: Objection.
d. THE COURT: Overruled.
e. D: Correct. There was a division of that nerve. I'm not sure I would say tore apart would be the word that I would use.
f. LP: Oh, there you go. You're getting a hint from your lawyer over here, so do you want to retract what you're saying?

The defendant was resisting LP's line of attack relatively well, but then made an error by agreeing to LP's loaded question, in which LP makes the proposition that is really at issue, that Dr. Tzeng was negligent, a presupposition by embedding it under a factive verb. This makes it difficult to answer for D the question in a straightforward way.

Since D had already repeatedly been asked about this issue, he wasn't paying attention. LP successfully attacks D's credibility in (6)f when D attempts to correct his mistake with (6)e, by seizing on a weakly implicated discourse connection between (6)c and (6)e of Result* (the commitment in (6)c caused the commitment in (6)e).

These examples suggest two general methods of deception: moves that implicate propositions that can't be committed to explicitly for strategic reasons, and moves that trap agents into making commitments they should rationally refrain from.

Another feature of attacks is that generally they work gradually in damaging an opponent's credibility. Perhaps no one move succeeds on its own in convincing the jury that the opponent is duplicitous or incompetent; rather a series of moves gradually move a jury to a skeptical view of the opponent over the course of a conversation. The victory conditions for our players are to succeed in eventually moving the jury to a position in which the opponent is no longer credible.

4 Dialogue model

We need a dialogue model in order to analyze our examples and attacks on credibility in more detail. We've already seen that we need to model as part of a speaker's contribution not only its compositional semantics but also its illocutionary effects, in particular the implicit discourse links between utterances, as these can trigger or convey attacks on credibility. We will therefore build on [15], as SDRT already offers a formal, logic-based approach of dialogue content (semantics + illocutionary effects).

[15] models the semantics of dialogue by assigning to each conversational agents a *commitment slate*. Each commitment slate contains a list of propositions that an agent is committed to, which involve rhetorical relations as well as elementary propositions. [15] model explicit and implicit agreements and denials of one agent about another agent's commitments. However, the analysis of credibility threats requires that we go a step further. Conversational agents explicitly or implicitly refer to, and dispute, others' commitments. They attack their opponent's credibility by exposing inconsistencies in something they claim the opponent committed to or implicated, and defend against such attacks by denying a commitment to content that the opponent claims they committed to or implicated. We need to represent the commitments of all speakers from their own and their interlocutors' points of view, as in [18]. Moreover, we need to represent arbitrary nesting of commitments explicitly. Recall example (2). In (2)b A corrects C's prior utterance, and thereby commits that C is committed to a false proposition p (N. is not coming because the meeting is cancelled). C rejects A's correction. But what C rejects is not the proposition that corrects p, but A's commitment that C commited to p. Therefore, C also commits that A commits that C commits that p. Further, we need to distinguish between weak and strong commitments: when an agent tries to misdirect another, he might for instance give a weak commitment the look of a stronger one. Thus our dialogue model will add three things to [15]: explicit nested commitments, the commitments of each agent from every agents' point of view and explicit strong and weak commitments.

Conversations proceed as follows in our model: speakers alternate turns, each performing a sequence of discourse moves. Because we are interested in commitments and attacks, we will not import the full machinery of SDRT here. We will symbolize clausal contents within a propositional language, but incorporate labels for speech acts and discourse relations so that we can roughly express discourse-structures following [1]. Crucially, however, our language allows us to embed discourse structures under 3 modal operators $[\ ]$, $\langle\ \rangle$ and N. A discourse move for an agent i is defined as a discourse-level proposition labelled by a speech act identifier. A discourse-level proposition is either a base-level proposition, a formula expressing commitment over a discourse structure (*i.e.* i commits that a label have some particular content), or a complex formula $R(\pi_1, \pi_2)$ where R is a coherence-relation symbol and π_1 and π_2 are speech act labels. A complex formula recursively involves previously introduced speech acts labels. The modalities make the language more expressive, since we can express commitments of different agents to different contents for a single speech-act. The formula $[\pi : \gamma]_i$ states that agent i commits that the content of the speech act π is γ. Hence, she also commits that the speaker of π commits to the discourse proposition γ. Its dual, $\langle\pi : \gamma\rangle_i$, expresses the proposition that it is possible for i that the content of π is γ. $N_i\varphi$ means that i defeasibly commits to the contents of the for-

mula φ. These modal operators express commitments over *discourse structures*. From this we retrieve commitments over *informational content* by looking at the content assigned to labels which are maximal for a given speaker. (labels that are not in the scope of another label of the same speaker): a speaker is committed to a content φ iff she commits the content of one of his maximal labels to be a proposition that entails φ.

Assume a set Φ of base-level propositions, a countably-infinite set of labels Π, a finite set of relation symbols $\mathcal{R}$ and a set of conversational agents I. In order to keep track of which agent x perform which speech act π, we assume Π partitioned in $|I|$ disjoint subsets $(\Pi_i)_{i\in I}$. We define $\text{spk}(\pi) =$ the unique $i \in X$ such that $\pi \in \Pi_i$.

$$\Gamma(\Phi) := \varphi \mid R(\pi_1, \pi_2) \mid [\delta]_i \mid N_i(\delta) \mid \langle\delta\rangle_i \mid \neg\gamma$$

$$\Delta := \pi : \gamma \mid \pi :?(\gamma) \mid \delta_1 \wedge \delta_2$$

Where the γ_i and δ_i respectively range over $\Gamma(\Phi)$ and Δ, the π_i and φ_i respectively range over Φ and Π, and i and R respectively range over I and $\mathcal{R}$.

Definition 1 (Model)**.** *A model $\mathcal{M}$ is a tuple $\langle W, v, (\triangleright_x)_{x\in X}, <\rangle$, where W denotes set of possible worlds, $v : \Phi \mapsto \wp(W)$ a coloration, $<$: $W \to W^2$ a function from worlds to partial orderings over W, and for each agent x, $\triangleright_x \subseteq W \times W$ is a transitive and euclidean accessibility relation.*

Our language has a dynamic semantics: the interpretation of a formula is context-change potential *i.e.* a relation between world-assigment pairs (w, σ). To account for polar question in our examples, we adopt a simplistic version of [13] and take propositions to semantically denote a set of set of worlds (a proposition denotes a set of possibilities which is partitioned into equivalence classes raised by questions). For intance, the question *whether p?* partitions a set of world in two, those worlds at which p on the one hand, and those at which $\neg p$ on the other. An assigment $\sigma : \Pi \times W \mapsto \wp(\wp(W))$ is a function that assigns a proposition as a set of set of worlds to a speech act label at a particular world. $\sigma(w, \pi)$ is roughly the (partitioned) set of worlds in which the interpretation of π at world w is true. Given a model $\mathcal{M}$, the function $[\![\cdot]\!]_{\mathcal{M}}$ maps each formula δ of the language to a binary relation $[\![\delta]\!]_{\mathcal{M}}$ over world-assignment pairs. Discourse-level assertoric propositions in $\Gamma(\varphi)$ always leave the assignment component unchanged and act as filters that let through only the worlds at which the proposition is true. Discourse moves in Δ on the other end modify the assignment. Another bit of needed machinery is for interpreting discourse relations. In our semantics each relation affects the contents assigned to its terms. Veridical relations like Explanation or Result will simply update the contextually given values to its terms with the semantic effects of the relation on those terms [1]. Non veridical relations like Correction or alternation place constraints on the truth of the contents associated with the terms at worlds verifying the relation in question. We need some notation first: assume a model $\mathcal{M} = \langle W, v, s, (\triangleright_x)_{x\in X}, <\rangle$. Let p denote a dynamic proposition (*i.e* a relation between world/assigments pairs). Define $|p|^\sigma_{\mathcal{M}}$ as $\{w \in W \mid (\sigma, w)\ p\ (\sigma, w)\}$ and $|?p|^\sigma_{\mathcal{M}}$ as $\{|p|^\sigma_{\mathcal{M}}, W \setminus |p|^\sigma_{\mathcal{M}}\}$. Define $Acc(w)$ as the set of set of world containing a single element which is the set of all worlds accessible from w: $Acc_x(w) = \{\{w' \mid w \triangleright_x w'\}\}$. Finally define the update operation $\star : \wp(\wp(W)) \times \wp(\wp(W)) \mapsto \wp(\wp(W))$ as $a \star b = \{x \cap y \colon x \in a \wedge y \in b\}$.

Definition 2 (Semantics)**.** *Discourse propositions:*

$$(w, \sigma)[\![\varphi]\!]_{\mathcal{M}}(w', \sigma') \text{ iff } \begin{cases} (\sigma, w) = (\sigma', w') \\ w \in v(\varphi) \end{cases}$$

$(w, \sigma)[\![R(\pi_1, \pi_2)]\!]_{\mathcal{M}}(w', \sigma')$ iff $(\sigma, w) = (\sigma', w')$ and $w \in I_R(\sigma(\pi_1, w), \sigma(\pi_2, w))$

$(\sigma, w)\ [\![[\delta]_x]\!]_{\mathcal{M}}\ (\sigma', w')$ iff $w = w'$ and $\forall w''\ w \triangleright_x w'' \to (\sigma, w'')[\![\delta]\!]_{\mathcal{M}}(\sigma', w'')$

$(\sigma, w)\ [\![\langle\delta\rangle_x]\!]_{\mathcal{M}}\ (\sigma', w')$ iff $w = w'$ and $\exists w''\ w \triangleright_x w'' \wedge (\sigma, w'')[\![\delta]\!]_{\mathcal{M}}(\sigma', w'')$

$(\sigma, w)\ [\![N_x\delta]\!]_{\mathcal{M}}\ (\sigma', w')$ iff $w = w'$ and $\forall u\ (w \triangleright_x u \wedge \forall v(w \triangleright_x v \to u \geq_w v)) \to (\sigma, u)\ [\![\delta]\!]_{\mathcal{M}}\ (\sigma', u)$

Discourse moves:

$(\sigma, w)\ [\![\pi : \gamma]\!]_{\mathcal{M}}\ (\sigma', w')$ iff $w = w'$ and $\sigma'(\pi, w) = \sigma(\pi, w) \star |\gamma|^\sigma_{\mathcal{M}} \star Acc_{\text{spk}(\pi)}(w)$

$(\sigma, w)\ [\![\delta_1 \wedge \delta_2]\!]_{\mathcal{M}}\ (\sigma', w')$ iff $w = w'$ and $(\sigma, w)\ [\![\delta_1]\!]_{\mathcal{M}} \circ [\![\delta_1]\!]_{\mathcal{M}}\ (\sigma', w')$

Armed with this semantics for formulas, we can now define the commitments of each agent i at every initial prefix (sequence of turns) in the conversation. Because commitments will depend on discourse structure, we define commitments at *maximal* labels in the logical forms for the turns (those that are not within the scope of any other label). Given a logical form for n conversational turns (or the whole conversation), we can define the commitment of the players:

Definition 3. $(\sigma, w)\ [\![C_i\varphi]\!]_{\mathcal{M}}\ (\sigma, w)$ *iff*

$$\exists\pi(\text{spk}(\pi, i) \wedge\ maximal(\pi) \\ \wedge \forall u(w \triangleright_x u \rightarrow \sigma(\pi, v) \subseteq |\varphi|))$$

We thus have a dynamic picture of how speakers' commitments evolve throughout a conversation.

Examples revisited. We start with example (2). [15] would analyse the two first turns as in table 1

turn	C's SDRS	A's SDRS
(2-a)	$\pi_1 : \neg N$ $\pi_2 : ccl\_meeting$ $\pi_3 : Res(\pi_1, \pi_2)$	
(2-b)		$\pi_4 : \neg Exp(\pi_1, \pi_2)$ $\pi_5 : Corr(\pi_3, \pi_4)$

Table 1: Analyisis of (2) following [15].

This is problematic, since A is committed to an absurdity. The semantic conditions of $Corr(\pi_3, \pi_4)$ require that the content of π_3 implies the negation of π_4, but $Res(\pi_1, \pi_2)$ does not imply $Exp(\pi_1, \pi_2)$ (the two are even contradictory). Keeping with the same kind of tabular representation as [15] our proposal amounts to further divide each cell of the table above in two, introducing A' interpretation of $C's$ moves, and repeating this process potentially infinitely to express arbitrary nestings as in table 2. For readability, we simplify the table by recopying at each step only the moves whose interpretation is controversial in the nested cells. In our language (2) is analysed as:

$$[\pi_1 : \neg N]_c \wedge [\pi_2 : ccl\_meeting]_c \\ \wedge [\pi_3 : Res(\pi_1, \pi_2)]_c \\ \wedge [\pi_4 : \neg Exp(\pi_1, \pi_3)]_a \wedge [\pi_5 : \langle \pi_3 : Exp(\pi_1, \pi_2)\rangle_c \\ \wedge \pi_5 : Corr(\pi_3, \pi_4)]_a \\ \wedge [[\pi_5 : \langle \pi_3 : Exp(\pi_1, \pi_2)\rangle_c]_a \\ \wedge \pi_6 : \neg C_x(Exp(\pi_1, \pi_2)) \wedge \pi_7 : Corr(\pi_5, \pi_6)]_c$$

Correcting move like π_5 triggers presupositions: here, a presuposition that c's move π_3 possibly commits him to the negation of π_4's content, accomodated as part of π_5's content. In the tabular representation, C's final move is:

C's SDRS		
C	A	
$\pi_6 :$	C	A
$\neg C_x(Ex(\pi_1, \pi_2))$	$\pi_3 :$	
$\pi_7 : Cor(\pi_5, \pi_6)$	$Exp(\pi_1, \pi_2)$	

In (2), we have only encountered explicit commitments $[\varphi]_x$. But in (1)b, A takes C's commitments to involve two possibilities, and he does not know which C has in fact committed to. Thus, in (1)b, A represents C's commitments as

$$[\pi_1 : \neg N]_c \wedge [\pi_2 : ccl\_meeting]_c \\ \wedge [\pi_3 : Res(\pi_1, \pi_2)]_c \\ \wedge [\pi_5 : \textit{Clar-q}(\pi_4, \pi_3)]_a \wedge [\pi_5 : \langle \pi_3 : Exp(\pi_1, \pi_2)\rangle_c \\ \wedge \pi_5 : \langle \pi_3 : Res(\pi_1, \pi_2)\rangle_c]_a$$

In (3), Bentsen (B) seizes on a weak implicature of Quayle's (Q). Q explicitly commits to a direct comparison between his experience in government and that of the young JFK, but B corrects a more general equivalence between the presidential promise of JFK and his own. If we symbolize the latter with JFK, we take B's turn to yield $[\pi' : \neg \texttt{JFK} \wedge \langle \pi : \texttt{JFK}\rangle_q \wedge \pi_2 : Cor(\pi, \pi')]_b$. We see that even a weak implicature is sufficient to warrant B's corrective move in π_2. The success of this attack relies on the jury's decision on the admissibility of $\langle \pi : \texttt{JFK}\rangle_q$, *i.e.* the possibility of a commitent of q to JFK. Finally, in (6), we see that LP commits that D is committed to a discourse link between the defense attorney and his own self-correction:

$$[[\pi_1 :?p]_{lp} \wedge \pi_2 : p \quad \wedge \pi_3 : QAP(\pi_1, \pi_2)]_d \\ \wedge [\pi_5 : Obj(\pi_4, \pi_3) \wedge \pi_9 : \langle \pi_7 : Res(\pi_5, \pi_6)\rangle_d \\ \wedge \pi_9 : [\pi_8 : Cor(\pi_3, \pi_6)\rangle_d]_{lp}$$

C's SDRS	A's SDRS	
$\pi_1 : \neg N$		
$\pi_2 : ccl\_meeting$		
$\pi_3 : Res(\pi_1, \pi_2)$		
	C	A
		$\pi_4 :$
	$\pi_3 :$	$\neg Exp(\pi_1, \pi_2)$
	$Exp(\pi_1, \pi_2)$	$\pi_5 :$
		$Corr(\pi_3, \pi_4)$

Table 2: Adding nested commitments

5 The strategic model

Speakers choose the sequences of discourse moves they do because they want to convey commitments that will make them look good in the eyes of the Jury; they also want to make an opponent look bad if possible by attacking her weak points. We call this their winning condition. We will assume as in [3] that speakers may have incomplete knowledge of the other players' moves, leading to nasty surprises as in (3), where Quayle clearly didn't anticipate Bentsen's move in (3)b. A final desirable feature of the strategic model is that the moves open to a participant that lead her to her winning condition may decrease or even vanish if her credibility is repeatedly attacked. Thus the underlying framework of a sequential game is essential for analyzing conversation.

During play, a player has to weigh whether to make a move that makes her look good but that is risky in that it can be attacked; if the attack has no grounded rebuttal [9], the move could be disastrous. Further, when an opponent j makes a move involving an implicature it is up to the player i to decide whether it can be taken as a safe commitment in the sense of [2], and to exploit it in subsequent conversational moves, as the prosecutor of (4) does; and conversely j has to weigh whether the player will take the implicature on board or not, as one of i's commitments. If not, the deceptive move may fail if an opponent makes a request for an explicit version of an implicated commitment as in (5).

All of these calculations depend on the effect of play on the Jury, who ultimately decides the winner according to positive points and lack of bad moves (inconsistencies or deceptions) on the part of i and other players. Our Jury entertains a space of possibilities concerning player types for the players and a probability distribution P over them. Our model is simple; we assume just two types for each player GOOD and BAD. At the start of the conversation the Jury entertains only the possibility that all players are GOOD; that is the probability distribution is such that $P(\text{BAD}_i) = 1 - P(\text{GOOD}_i) = 0$ for any player i. As the conversation proceeds, $P(\text{BAD}_i)$ is successively updated given what has happened over the last turn; i.e. $P_n(\text{BAD}_i) = P_{n-1}(\text{BAD}_i/t_n)$. As long as the opponent does not convincingly refute the arguments of i at n, $P_n(\text{BAD}_i) = P_{n-1}(\text{BAD}_i)$. However, a successful attack on, say, i by j at turn t_n, which results in a refutation of an argument by i with no convincing rebuttal gets the Jury to update P via Bayesian conditionalisation such that $P_n(\text{BAD}_i) > P_{n-1}(\text{BAD}_i)$.

The effect of a higher probability on BAD_i is that the positive reasons advanced by i are given a lower score; that is the effect of a bad reputation—the good things you say get discounted. Thus, if the positive arguments by i in her favor provide some positive score σ^i, then the effect on the Jury at turn n is:[1]

$$\text{overall-score}^i_n = P_n(\text{GOOD}_i)(\sigma^i_n) \qquad (1)$$

Our model should also reflect the duplicitous nature of weak implicatures that agents don't dare put out as full commitments. So the update of the probability on $P_n(\text{BAD}_i)$ will depend on (i) the strength of the implicature, (ii) whether the attack is successful in so far as there is no rebuttal that refutes it. For weak implicatures, there is a rebuttal: you misinterpreted what I said; but it renders the move useless for the player. The upshot of our model is that agents pay dearly if their credibility is successfully attacked when they advance a weak implicature, as evidenced in the example (3). What exactly was Quayle's (DQ) mistake? It was that he weakly implicated that he was of the same caliber as JFK, and it is this implicature that Bentsen (B) seizes on and shows to be ridiculous. He also implicates that DQ insinuated the direct comparison without directly saying so, which is a deceptive move. B's attack was very successful, especially since DQ did not vigorously rejoin with *you misinterpreted me unfairly*. Our model considerably increases $P_n(\text{BAD}_i)$ conditional on B's attack, rendering DQ's successful points much weaker. Not only did B refute DQ's argument and expose his deceptive move, but he affected the overall outcome of the debate.

In example (4) on the other hand Bronston (B) carries off his strongly implicated commitment to an answer in (4)d without being challenged by the prosecutor (P). However, it was somewhat dangerous. Had P continued as in (5), B would have had to contradict $N_B(\neg\text{bank account})$ with $[\text{bank account}]_B$, the two modal formulas being inconsistent. Bronston could have claimed that he had not understood the first prosecutor's question as being directly about him, but the response

[1] In future work, we plan to experiment with refinements of this basic idea, such as using updated probabilities to score continuations.

would have been weak and our intuition is that his credibility would have suffered. Example (6) exhibits a slightly different pattern: D has committed to there being no negligence in the operation, but in (6)b, he commits to the presupposition of the question that entails negligence. Conditional on such a contradiction, $P(\text{BAD}_R)$ increases, but not much because it was a trick question. But when D attempts to retract his affirmative answer to LP's biased question, then pins on him reasoning that attacks his credibility as an impartial witness via a weakly implicated connection between LD's and D's contributions. At this point, $P(\text{BAD}_D)$ increases considerably, and weakens D's testimony in the eyes of the Jury.

Our model predicts that as a player's credibility is repeatedly attacked and duplicitous moves are exposed, her credibility decreases monotonically. As a consequence, after a certain point the player may have no moves open to her that achieve her winning condition—the probability of her being of BAD type is now too high.

6 Related work

Our work assumes a commitment based view of conversation rather than one based on the internal, mental states of the participants [14, 20, 19] and builds on and complements the model proposed in [18], which in turn extends [7]. They introduce a dynamic Bayesian model for discourse actions based on prior moves. Our paper is more limited in scope but also goes into more detail: our model details how attacks on credibility function with respect to various types of commitments that come from different kinds of discourse moves; we show that even in simple conversations levels of embedded commitments can be very complex (contrary to a suggestion of [18]); and our Bayesian update on player types details a part of the picture of [18].

Our model assumes a sequential game view of conversation, differing from extended signaling games [4], and uses a notion of credibility which differs from the standard one in signaling games, according to which a message is credible iff its standardly accepted content understood as a set of evaluation points is a superset of its meaning in reflective equilibrium (roughly how the message content affects belief). In strategic environments of the sort we have in mind, signaling games have severe limitations: in our strategic contexts, a player will send a message only if it benefits him, but then that message will not benefit his opponent. In a signaling game, the opponent should rationally ignore it [6]. However, in a debate, it would be irrational to ignore the message of the opponent. Our notion of credibility does not mix belief and action in the way signaling games do, and is immune to this problem. A further problem with signaling games is that they assume common knowledge of the preferences of each player over moves. But if these are used to define or to guide credibility, then there is no room for maneuver or deception, which is manifest in our examples. However, our model leaves a place for a signaling game analysis between the Jury and the players, which we will pursue in future work.

Related to our work are also recent attempts to investigate argumentation in actual dialogue [5]. Argumentation theory provides a framework for analyzing attacks and counterattacks [9]. We have given much more linguistic detail on how such attacks are carried out and how this can affect ones' strategy in conversation. On the other hand, we have presented a general model for credibility in strategic conversation. Different contexts may affect the parameters of the model that we have set up. For instance,[2] sometimes the Jury may be a participant in the conversation in the sense that it is allowed to ask questions, sometimes not. Given a particular context, Jury might also function according to persuasion rules that are different from the simple one we have used in section 5. We have chosen simple settings to illustrate our model. Finally, we have not gone into the details of how particular conversational contexts may dictate specific linguistic forms of attacks and defense, e.g., [8]. Our model is general enough, we believe, so that we can tune the parameters to fit the particularities of specific contexts.

7 Conclusions

In this paper, we have presented new notion of credibility and attacks on credibility that are relevant to conversations in strategic settings where interlocutor preferences may be opposed. We have developed a dialogue model extending both [15] and [18] with a semantics for dialogue turns and commitments that allows for arbitrary nestings of commitments. We have also shown that this complexity is required to analyze many examples of dialogue with attacks on credibility.

[2] Thanks to a Semdial reviewer for this point.

References

[1] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

[2] N. Asher and A. Lascarides. Strategic conversation. *Semantics and Pragmatics*, Vol 6.2:1–62, 2013.

[3] N. Asher and S. Paul. Infinite games with uncertain moves. In F. Mogavero, A. Murano, and M. Vardi, editors, *Proceedings of the First Workshop on Strategic Reasoning, ET-PCS 2013, Rome, Italy*, pages 25–32, Rome, Italy, 2013. Springer.

[4] R. Aumann and S. Hart. Long cheap talk. *Econometrica*, 71(6):1619–1660, 2003.

[5] Elena Cabrio, Sara Tonelli, and Serena Villata. A Natural Language Account for Argumentation Schemes. In *AI*IA - XIII Conference of the Italian Association for Artificial Intelligence - 2013*, Turin, Italie, December 2013. Springer.

[6] V. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

[7] D. DeVault and M. Stone. Managing ambiguities across utterances in dialogue. In *Proceedings from the International Workshop on the Semantics and Pragmatics of Dialogue (DECALOG 2007)*, Trento, Italy, 2007.

[8] Paul Drew. Contested evidence in courtroom cross-examination: The case of a trial for rape. *Talk at work: Interaction in institutional settings*, pages 470–520, 1992.

[9] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[10] Joseph Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5(4):514–531, 1993.

[11] M. Franke, T. de Jager, and R. van Rooij. Relevance in cooperation and conflict. *Journal of Logic and Language*, 2009.

[12] R. Friedman and P. Malone. *Rules of the Road: A Plaintiff Lawyers Guide to Proving Liability*. Trial Guides, 2nd edition, 2010.

[13] J. Groenendijk and M. Stokhof. *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, Centrale Interfaculteit, Amsterdam, 1984.

[14] C. Hamblin. *Imperatives*. Blackwells, 1987.

[15] A. Lascarides and N. Asher. Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158, 2009.

[16] M. Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, Department of Computer Science, King's College, London, 2004.

[17] L.M. Solan and P.M. Tiersma. *Speaking of Crime: The Language of Criminal Justice*. University of Chicago Press, Chicago, IL, 2005.

[18] M. Stone and A. Lascarides. Grounding as implicature. In *Proceedings of the 14th SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue*, pages 51–58, Poznan, 2010.

[19] D. Traum. Computational models of non-cooperative dialogue. In *Proceedings of the International Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL)*, London, 2008.

[20] D. Traum and J. Allen. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL94)*, pages 1–8, Las Cruces, New Mexico, 1994.

Poster Abstracts

MILLA – Multimodal Interactive Language Learning Agent

João Paulo Cabral[1], Nick Campbell[1], Shree Ganesh[2], Emer Gilmartin[1], Fasih Haider[1], Eamonn Kenny[1], Mina Kheirkhah[3], Andrew Murphy[1], Neasa Ní Chiaráin[1], Thomas Pellegrini[4], Odei Rey Orozko[5]

Trinity College Dublin, Ireland[1]; GCDH-University of Goettingen, Germany[2]; Institute for Advanced Studies in Basic Sciences, Zanjan, Iran[3]; Université de Toulouse ; IRIT, France[4]; Universidad del País Vasco, Bilbao, Spain[5]

1 Background

Learning a new language involves the acquisition and integration of a range of skills. A human tutor aids learners by (i) providing tasks suitable to the learner's needs, (ii) monitoring progress and adapting task content and delivery style, and (iii) providing a source of speaking practice and motivation. With the advent of audiovisual technology and the communicative paradigm in language pedagogy, focus has shifted from written grammar and translation to developing communicative competence in listening and spoken production. The Common European Framework of Reference for Language Learning and Teaching (CEFR) recently added a more integrative fifth skill – spoken interaction - to the traditional four skills – reading and listening, and writing and speaking (Little, 2006) . While second languages have always been learned conversationally with negotiation of meaning between speakers of different languages sharing living or working environments, these methods did not figure in formal (funded) settings. However, with increased mobility and globalisation, many learners now need language as a practical tool rather than simply as an academic achievement (Gilmartin, 2008). Developments in Computer Assisted Language Learning (CALL) have resulted in free and commercial language learning material for autonomous study. Much of this material transfers well-established text and audiovisual exercises to the computer screen. These resources greatly help develop discrete skills, but the challenge of providing tuition and practice in the 'fifth skill', spoken interaction, remains. MILLA, developed at the 2014 eNTERFACE workshop in Bilbao is a multimodal spoken dialogue system combining custom modules with existing web resources in a balanced curriculum, and, by integrating spoken dialogue, modelling some of the advantages of a human tutor.

2 MILLA System Components

Tuition Manager: MILLA's spoken dialogue Tuition Manager (Figure 1) consults a two-level curriculum of language learning tasks, a learner record and learner state module to greet and enroll learners, offer language learning submodules, provide feedback, and monitor user state with Kinect sensors. All of the tuition manager's interaction with the user can be performed using speech through a Cereproc Text-to-Speech (TTS) voice and Cereproc's Python SDK (Cereproc, 2014), and understanding via CMU's Sphinx4 ASR (Walker et al., 2004) through custom Python bindings using W3C compliant Java Speech Format Grammars.

Tasks include spoken dialogue practice with two chatbots, first language (L1) focused and general pronunciation, and grammar and vocabulary exercises. Several speech recognition (ASR) engines (HTK, Google Speech) and text-to speech (TTS) voices (Mac and Windows system voices, Google Speech) are used in the modules to meet the demands of particular tasks and to provide a cast of voice characters which provide a variety of speech models to the learner. Microsoft's Kinect SDK ('Kinect for Windows SDK', 2014) is used for gesture recognition and as a platform for affect recognition. The tuition manager and all interfaces are written in Python 2.7, with additional C#, Javascript, Java, and Bash coding in the Kinect, chat, Sphinx4, and pronunciation elements. For rapid prototyping the dialogue modules were first written in VoiceXML, then ported to Python modules.

Pronunciation Tuition: MILLA incorporates two pronunciation modules, based on comparison of learner production with model production using the Goodness of Pronunciation (GOP) algorithm (Witt & Young, 2000). GOP scoring involves two phases: 1) a free phone loop recognition phase which determines the most likely phone sequence given the input speech without

giving the ASR any information about the target sentence, and 2) a forced alignment phase which provides the ASR with the orthographic transcription and force aligns the speech signal with the expected phone sequence. Comparison of the log-likelihoods of the forced alignment and free recognition phases produces a GOP score.

The first module is a focused pronunciation tutor using HTK ASR with the five-state 32 Gaussian mixture monophone acoustic models provided with the Penn Aligner toolkit (Young, n.d.; Yuan & Liberman, 2008) on the system's local machine. In this module, phone specific threshold scores were set by artificially inserting errors in the pronunciation lexicon and running the algorithm on native recordings, as in (Kanters, Cucchiarini, & Strik, 2009). After preliminary tests, we constrained the free phone loop recogniser for more robust behavior, using phone confusions common in specific L1's to define constrained phone grammars. A database of common errors in several L1s with test utterances was built into the curriculum.

The second module, MySpeech, is a phrase level trainer hosted on University College Dublin's cluster and accessed by the system via Internet (Cabral et al., 2012). It tests pronunciation at several difficulty levels as described in (Kane & Carson-Berndsen, 2011). Difficulty levels are introduced by incorporating Broad Phonetic Groups (BPGs) to cluster similar phones. A BFG consists of phones that share similar articulatory feature information, for example plosives and fricatives. There are three difficulty levels in the MySpeech system: easy, medium and hard – the easiest level includes a greater number of BPGs in comparison to the harder levels. The MySpeech web interface consists of several numbered panels for the users to select sentences and practice their pronunciation by listening to the selected sentence spoken by a native speaker and record their own version of the same sentence. Finally, the results panel shows the detected mispronunciation errors of a submitted utterance using darker colours.

Spoken Interaction Tuition (Chat): To provide spoken interaction practice, MILLA sends the user to Michael (Level1) or Susan (Level 2), two chatbots created using the Pandorabots web-based chatbot hosting service (Wallace, 2003). The bots were first implemented in text-to-text form in AIML (Artificial Intelligence Markup Language) and then TTS and ASR were added through the Web Speech API, conforming to W3C standards (W3C, 2014). Based on consultation with language teachers and learners, the system allows users to speak to the bot, or type chat responses. A chat log was also implemented in the interface, allowing the user to read back or replay previous interactions.

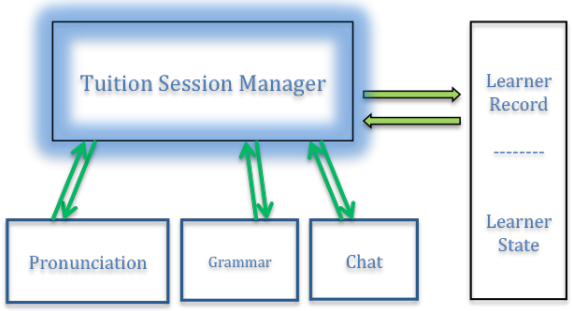


Figure 1 MILLA Overview

Grammar, Vocabulary and External Resources: MILLA's curriculum includes a number of graded activities from the OUP's English File and the British Council's Learn English websites. Wherever possible the system scrapes any scores returned for exercises and incorporates them into the learner's record, while in other cases the progression and scoring system includes a time required to be spent on the exercises before the user progresses to the next exercises. There are also custom morphology and syntax exercises created using Voxeo Prophecy to be ported to MILLA.

User State and Gesture Recognition: MILLA includes a learner state module to eventually infer learner boredom or involvement. As a first pass, gestures indicating various commands were designed and incorporated into the system using Microsoft's Kinect SDK. The current implementation comprises four gestures (Stop, I don't know, Swipe Left/Right), which were designed by tracking the skeletal movements involved and extracting joint coordinates on the x, y, and z planes to train the recognition process. Python's socket programming modules were used to communicate between the Windows machine running the Kinect and the Mac laptop hosting MILLA.

3 Future work

MILLA is an ongoing project. In particular, work is in progress to add a Graphical User Interface and avatar to provide a more immersive version and several new modules are planned. User trials are planned for the academic year 2014-15 in several centres providing language training to immigrants in Ireland.

Acknowledgments

João Paulo Cabral and Eamonn Kenny are supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie) at Trinity College Dublin.Sree Ganesh is supported by GCDH-University of Goettingen. Emer Gilmartin is supported by the Science Foundation Ireland Fastnet Project, grant 09/IN.1/I2631. Neasa Ní Chiaráin and Andrew Murphy are supported by the ABAIR Project, funded by the Irish Government's Department of Arts, Heritage, and the Gaeltacht.

References

Cabral, J. P., Kane, M., Ahmed, Z., Abou-Zleikha, M., Székely, E., Zahra, A., … Schlögl, S. (2012). Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz. In *LREC* (pp. 4136–4142).

CereVoice Engine Text-to-Speech SDK | CereProc Text-to-Speech. (2014). Retrieved 7 July 2014, from https://www.cereproc.com/en/products/sdk

Gilmartin, E. (2008). Language Training for Adult Refugees: The Integrate Ireland Experience. *Adult Learner: The Irish Journal of Adult and Community Education*, *97*, 110.

Kane, M., & Carson-Berndsen, J. (2011). Multiple source phoneme recognition aided by articulatory features. In *Modern Approaches in Applied Intelligence* (pp. 426–435). Springer.

Kanters, S., Cucchiarini, C., & Strik, H. (2009). The goodness of pronunciation algorithm: a detailed performance study. In *SLaTE* (pp. 49–52).

Kinect for Windows SDK. (2014). Retrieved 7 July 2014, from http://msdn.microsoft.com/en-us/library/hh855347.aspx

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, *39*(03), 167–190.

W3C. (2014). *Web Speech API Specification*. Retrieved 7 July 2014, from https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., … Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.

Wallace, R. S. (2003). *Be Your Own Botmaster: The Step By Step Guide to Creating, Hosting and Selling Your Own AI Chat Bot On Pandorabots*. ALICE AI foundations, Incorporated.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2), 95–108.

Young, S. (n.d.). *HTK Speech Recognition Toolkit*. Retrieved 7 July 2014, from http://htk.eng.cam.ac.uk/

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*(5), 3878.

Dialogue Structure of Coaching Sessions

Iwan de Kok[1,2], **Julian Hough**[1,3], **Cornelia Frank**[1,4],
David Schlangen[3], and **Stefan Kopp**[1,2]
[1]CITEC, [2]Social Cognitive Systems,
[3]Dialogue Systems Group, [4]Neurocognition and Action (Biomechanics)
Bielefeld University
idekok@techfak.uni-bielefeld.de

Abstract

We report initial findings of the ICSPACE ('Intelligent Coaching Space') project on virtual coaching. We describe the gathering of a corpus of dyadic squat coaching interactions and initial high-level models of the structure of these sessions.

1 Introduction

While interactive tutoring systems which perform factual teaching have been established for some time (Litman and Silliman, 2004; Graesser et al., 2005), dialogue systems capable of skill coaching are much rarer. We introduce preliminary work on the ICSPACE ('Intelligent Coaching Space') project, which aims to create a virtual intelligent coaching agent in an interactive environment to train users to perform complex motor actions.

Coaching physical movement skills requires combining communication with real-time tracking, assessing and correcting the motor action of the coachee. In particular, giving online feedback while the coachee is carrying out an exercise (Sigrist et al., 2013) is an interesting challenge imposing specific requirements on the system .

To identify these requirements more precisely we analyse two recordings of a professional coach training individuals to perform a squat. We focus on the overall dialogue structure and observe which dialogue situations arise.

2 Recordings

We invited a professional coach to our lab to record two coaching sessions. The coach was asked to instruct coachees how to do a squat as he would teach it in the gym. The coachees (one female (A), one male (B)) were familiar with doing squats, although they had not received instruction from a professional coach before. Each interaction lasted between 4 and 5 minutes.

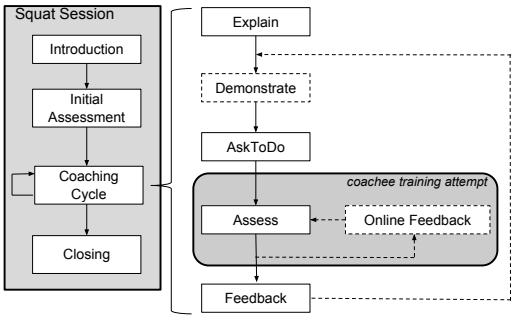


Figure 1: Overall structure of a squat coaching session (left) and structure of a coaching cycle within the session (right).

3 Dialogue Structure

In analysing the two dialogues, several commonalities to the structure of the interaction were observed, as well as some interesting differences. From these dialogues a common overall coaching structure can be inferred, represented in the left-hand shaded box in Figure 1 – temporal dependencies are represented by arrows.

In both dialogues, there was an *Introduction* phase where rapport with the coachee was established by the coach, consisting of questions about personal details and also establishing their previous experience of the squat exercise.

In the second phase of *Initial Assessment*, while in A's session this was completed after lengthy explanation, for B the coach begins by asking the coachee to do a squat before any explanation. In both cases the coach assesses the coachee's ability at performing the squat, identifying the sub-movements and aspects of their technique which fall short of the coachee's potential, and subsequently planning the following coaching behaviour.

What follows are a series of *coaching cycles*, each explaining a particular area in which the movement is being executed below the potential

of the coachee, or which the coach wants to bring to attention to ensure the coachee will continue executing that aspect correctly. As soon as the coach evaluates the action as being performed appropriately after several iterations, the session is *closed*.

Coaching cycles Each coaching cycle follows a similar structure as depicted in the right-hand diagram of Figure 1 – the optional components and transitions have dashed lines. In both dialogues, the coach starts by *explaining* the particular (aspect of the) movement this coaching cycle will focus on (see visually, top-left in Figure 2). Often the coach *demonstrates* the movement and highlights the area of interest with gestures. The coach then *asks* the coachee to perform the movement and they comply, else this is done without prompting as the coachee takes initiative (as was the case in the top-right image of Figure 2). The latter case shows the possibility of mixed initiative in coaching dialogues and is analogous to question accomodation in issue-based dialogue management (Larsson, 2002), however the "answer" to the accomodated question here is non-linguistic.

As the coachee attempts the squat, either in a single effort or in repetition, the coach *assesses* the movement being performed by the coachee (bottom-left in Figure 2) and may give *online feedback* during execution to adjust the movement, either in the form of short utterances as verbal feedback, gesture or even by performing the movement synchronously with the coachee (bottom-right in Figure 2). Once the coach is satisfied with the result or can not adjust the movement during execution, he will stop the coachee and give more lengthy final *feedback* on the movement. During this stage he will often explain why this particular (aspect of the) movement is important. If the coach is satisfied he moves on to the next aspect (if there are any remaining to be corrected), otherwise the cycle will be partially repeated by either another demonstration or request to try again.

Spatial positioning of the coach The coaching cycles were not only noticable in the dialogue structure, but also in the coach's movement. The explanation and feedback phases were always performed in front of the coachee – the *instruction space* (see top two images in Figure 2). During the assessment and online feedback phases the coach usually moved to the side of the coachee to get a good profile view– the *observation space*

Figure 2: Images from the recordings showing *explaining* (top-left), *demonstrating* with user initiated following of (top-right), *assessing* (bottom-left) and finally giving *online feedback* on (bottom-right) the squat movement. The images also highlight the *instruction space* (top) and *observation space* (bottom).

(see bottom two screenshots in Figure 2). Initial demonstrations are usually performed in the instruction space. If the coach demonstrates the movement during the online feedback phase, this is performed in the observation space. This multi-locational instruction behaviour is similar to that exhibited by music teachers in instrumental lessons, who tend to move between the *work zone* and the *listening zone* (Duffy and Healey, 2012).

4 Conclusion

We have analyzed the dialogue structure of two exemplary squat coaching sessions to identify the requirements of a virtual intelligent coaching system. The initial recordings show the need for multi-modal turn-taking, use of different spaces for different coaching phases and the ability to generate fast incremental feedback as described by (Kopp et al., 2013) during phases of *online feedback*. As a next step we plan to collect more recordings with different coaches with coachees of different skill levels. This should give evidence for whether the dialogue structure we hypothesize generalizes to the domain, and it will inform the design of an artificial coach capable of online feedback and dialogue in coaching.

Acknowledgments

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence EXC 277 in Cognitive Interaction Technology (CITEC).

References

Sam Duffy and Patrick GT Healey. 2012. Spatial coordination in music tuition. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, 48(4):612–618.

Stefan Kopp, Herwin van Welbergen, Ramin Yaghoubzadeh, and Hendrik Buschmeier. 2013. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces*, pages 1–12.

Staffan Larsson. 2002. *Issue-based dialogue management*. Department of Linguistics, Göteborg University.

Diane J Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration Papers at HLT-NAACL 2004*, pages 5–8. Association for Computational Linguistics.

Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. 2013. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic bulletin & review*, 20(1):21–53.

Getting to Know Users: Accounting for the Variability in User Ratings

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser and Oliver Lemon
Heriot-Watt University
Mathematical and Computer Sciences
Edinburgh EH14 4AS, UK
n.s.dethlefs@hw.ac.uk

Abstract

Evaluations of dialogue systems and language generators often rely on subjective user ratings to assess output quality and performance. Humans however vary in their preferences so that estimating an accurate prediction model is difficult. Using a method that clusters utterances based on their linguistic features and ratings (Dethlefs et al., 2014), we discuss the possibility of obtaining user feedback implicitly during an interaction. This approach promises better predictions of user preferences through continuous re-estimation.

1 Introduction

Given the subjective nature of human language, many evaluation studies in dialogue systems and natural language generation rely on subjective user ratings to assess performance and acceptability. A shared problem however is that humans vary considerably in their individual preferences, making it difficult to estimate an accurate prediction model. To account for individual preferences and still make accurate predictions, in Dethlefs et al. (2014) we proposed to cluster utterances based on their linguistic properties and the ratings they receive from groups of individual users. Results confirmed that prediction accuracy improves significantly in this way: predictive models based on clusters of ratings lead to significantly better predictions than models based on an average population of ratings–as is currently state of the art.

The required clusters can be obtained from minimal information about an individuals user's preferences, such as a single user rating alone. One drawback of our method so far, however, is that it remains unclear how user ratings can best be obtained during an ongoing human-computer interaction. Requesting ratings explicitly may be the easiest way, but can disrupt interactions. Here, we discuss alternatives based on (a) the interaction history, (b) interactive alignment, and (c) multimodal information. We discuss the potential of each of these ideas to implicitly elicit user feedback on system utterances during an interaction.

2 State of the Art

The problem of variability in subjective user ratings has been recognised by various authors in different domains such as recommender systems (O'Mahony et al., 2006; Amatriain et al., 2009), sentiment analysis (Pang and Lee, 2005), content selection (Jordan and Walker, 2005; Dale and Viethen, 2009) and surface realisation (Walker et al., 2007; Dethlefs et al., 2014). The primary method of capturing individual differences in statistical models so far has been to train separate models for individual users (Dale and Viethen, 2009; Walker et al., 2007). In practice, this can often be done by including the user's ID as a feature for classification or regression. This tends to significantly improve performance for the user in question, but fails to generalise to users with no prior ratings. We can therefore distinguish (a) systems that estimate prediction models from an average population of users–and thereby ignore the existing variability; and (b) systems that are trained for individual users and fail to generalise to unseen instances.

3 Using Clustering to Account for Variable User Ratings

In Dethlefs et al. (2014), we have presented an approach that aims to find a middle ground between making predictions from an average population of users and training an individual model for each new user. Figure 1 provides an illustration of the approach. In essence, the idea is to learn a mapping between the linguistic features of a group of utterances that receive similar ratings, e.g. ratings

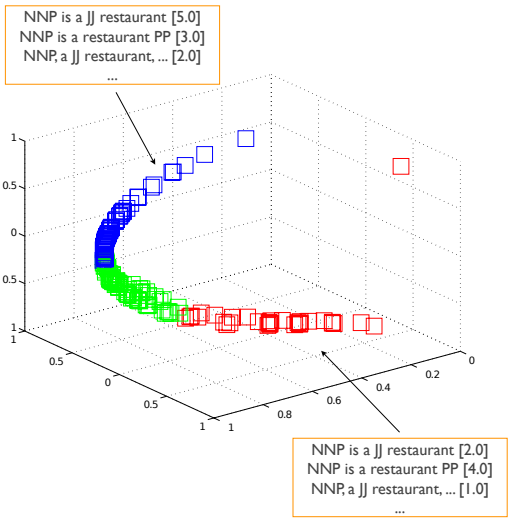


Figure 1: Clusters are estimated based on linguistic features and ratings. Prediction is then aided by estimating which cluster a new user might rate according to. Users in the same cluster (indicated in different colours) tend to rate utterances similarly.

for *politeness* on a scale of 1-5. We used multiple multivariate regression and features included lexical information, such as the presence of individual words, the average tf-idf score of an utterance, and syntactic features such as the depth of syntactic embedding. Clusters are identified from pair-wise similarities between data points using the Kullback-Leibler divergence (Cuayáhuitl et al., 2005). A spectral clustering algorithm performs dimensionality reduction and clusters similar pairs of linguistic features and user ratings into the same cluster and dissimilar pairs into separate clusters. Results have shown that minimal information on user preferences is sufficient to perform significantly better than based on an average population of users. Please see Dethlefs et al. (2014) for details on the approach and an evaluation.

4 Discussion

This section discusses three possible options of obtaining user feedback during an interaction.

Interaction Context including dialogue moves that follow a system utterance or incremental phenomena such as barge-ins or backchannels can all offer insights into a user's perception of an ongoing interaction (Janarthanam and Lemon, 2014). For example, barge-ins and unforeseen dialogue moves can be indicative of a problematic dialogue, whereas backchannelling and alignment with the system can indicate success. Based on this, a possibility is to extend the PARADISE framework (Walker et al., 1997) by estimating a regression model that predicts user ratings based on incremental dialogue phenomena in an online fashion. However, it is likely that such phenomena also exhibit variation between individual users. They can therefore provide feedback on subjective as well as objective evaluation scales.

Interactive Alignment could be applied under the hypothesis that adapting to the linguistic features found in users' speech would have a favourable influence on their perception of the system and lead to positive ratings. This assumption is based on psycholinguistic evidence that humans prefer to interact with humans that align with them (Levelt and Kelter, 1982). Further, computational studies have shown that interactive alignment in human-computer interaction can be created and recognised by users (Brockmann et al., 2005; Isard et al., 2006; Dethlefs, 2013). In our case, results of the ASR could be analysed and linguistic features extracted. An experimental study would have to confirm that such alignment is plausible, noticeable to users and perceived positively.

Multimodal Information could provide valuable feedback cues, including user hesitations and pauses or even gesture recognition or eye-tracking. Ultimately, our goal is to use non-verbal cues as feedback signals in an interaction so that system behaviour can be continuously re-estimated and improved (Cuayáhuitl and Dethlefs, 2011). Perceptive cues such as the user frowning, losing attention, or hesitating regarding the next step to take in the interaction could indicate problems in the interaction, while smiling or continued attention could be interpreted as positive cues. A data collection and analysis would need to explore the full range of multimodal cues available.

Future work will explore these ideas and analyse their practical advantages and drawbacks. To do this, we will use the PARLANCE system, a data-driven, incremental and spoken interactive system (Hastie et al., 2013), which also exists as a mobile app (Hastie et al., 2014). Implicit feedback elicitations could thus be combined with explicit feedback to gain more information on users and allow the personalisation of system output.

Acknowledgments

This research was funded by the EC FP7 programme FP7/2011-14 under grant agreement no. 287615 (PARLANCE).

References

Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. 2009. I like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In *In the 17th International Conference on User Modelling, Adaptation, and Personalisation (UMAP)*, pages 247–258, Trento, Italy. Springer-Verlag.

Carsten Brockmann, Amy Isard, Jon Oberlander, and Michael White. 2005. Modelling alignment for affective dialogue. In *Proceedings of the UM-05 Workshop on Adapting the Interaction Style to Affective Factors*.

Heriberto Cuayáhuitl and Nina Dethlefs. 2011. Optimizing Situated Dialogue Management in Unknown Environments. In *Proceedings of INTERSPEECH*, pages 1009–1012.

Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. 2005. Human-Computer Dialogue Simulation Using Hidden Markov Models. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, Puerto Rico.

Robert Dale and Jette Viethen. 2009. Referring Expression Generation Through Attribute-Based Heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, Athens, Greece.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based Prediction of User Ratings for Stylistic Surface Realisation. In *Proceedings of the European Chapter of the Annual Meeting of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden.

Nina Dethlefs. 2013. *Hierarchical Joint Learning for Natural Language Generation*. PhD Thesis, University of Bremen, Germany.

Helen Hastie, Marie-Aude Aufaure, Panos Alexopoulos, Heriberto Cuayhuitl, Nina Dethlefs, James Henderson Milica Gasic, Oliver Lemon, Xingkun Liu, Peter Mika, Nesrine Ben Mustapha, Verena Rieser, Blaise Thomson, Pirros Tsiakoulis, Yves Vanrompay, Boris Villazon-Terrazas, and Steve Young. 2013. Demonstration of the PARLANCE System: A Data-Driven, Incremental, Spoken Dialogue System for Interactive Search. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*.

Helen Hastie, Marie-Aude Aufaure, Panos Alexopoulos, Hughes Bouchard, Heriberto Cuayáhuitl, Nina Dethlefs, Milica Gasic, James Henderson, Oliver Lemon, Xingkun Liu, Peter Mika, Nesrine Ben Mustapha, Tim Potter, Verena Rieser, Blaise Thomson, Pirros Tsiakoulis, Yves Vanrompay, Boris Villa-Terrazas, Majid Yazdani, Steve Young, and Yanchao Yu. 2014. The PARLANCE Mobile App for Interactive Search in English and Mandarin. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*.

Amy Isard, Carsten Brockmann, and Jon Oberlander. 2006. Individuality and Alignment in Generated Dialogues. In *Proceedings of the 4th International Natural Language Generation Conference (INLG)*, Sydney, Australia.

Srini Janarthanam and Oliver Lemon. 2014. Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics*. (in press).

Pamela Jordan and Marilyn Walker. 2005. Learning Content Selection Rules for Generating Object Descriptions in Dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Willem Levelt and Stephanie Kelter. 1982. Surface form and memory in question answering. *Cognitive Psychology*, 14:78–106.

Michael O'Mahony, Neil Hurley, and Guénolé Silvestre. 2006. Detecting Noise in Recommender System Databases. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)s*. ACM Press.

Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marilyn Walker, Diane Litman, Candace Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, Madrid, Spain.

Marilyn Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and Domain Adaptation in Sentence Planning for Dialogue. *Journal of Artificial Intelligence Research*, 30(1):413–456.

Learning to manage risk in non-cooperative dialogues

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

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

Abstract

We investigate statistical dialogue agents which learn to perform non-cooperative dialogue moves in order to complete their own objectives in a stochastic trading game. We show that, when given the ability to perform both cooperative and non-cooperative dialogue moves, such an agent can learn to bluff and to lie so as to win games more often – against a variety of adversaries, and under various conditions such as risking penalties for being caught in deception. Here we present new results showing how learned non-cooperative dialogue strategies change depending on a) how severe the penalty is for being caught being non-cooperative, and b) how risky the non-cooperative behaviour is (i.e. the probability of being caught). For example, we show that a non-cooperative dialogue agent can learn to win an additional 4.5% of games against a strong rule-based adversary, even when there is an additional 10% chance of being caught (exposed) every time it attempts a non-cooperative (manipulative) move, when the penalty for being caught is that the adversary will no longer trade.

1 Introduction

Non-cooperative dialogues, where an agent may act to satisfy its own goals rather than those of other participants, are of practical and theoretical interest (Georgila and Traum, 2011), and the game-theoretic underpinnings of non-Gricean behaviour are actively being investigated (Asher and Lascarides, 2008). For example, it may be advantageous for an automated agent not to be fully cooperative when trying to gather information from a human, and when trying to persuade, argue, or debate, when trying to sell them something, when trying to detect illegal activity (for example on internet chat sites), or in the area of believable characters in video games and educational simulations (Georgila and Traum, 2011; Shim and Arkin, 2013). Another arena in which non-cooperative dialogue behaviour is desirable is in negotiation (Traum, 2008; Nouri and Traum, 2014), where hiding information (and even outright lying) can be advantageous. Indeed, Dennett argues that deception capability is required for higher-order intentionality in AI (Dennett, 1997).

A complementary research direction in recent years has been the use of machine learning methods to automatically optimise *cooperative* dialogue management - i.e. the decision of what dialogue move to make next in a conversation, in order to maximise an agent's overall long-term expected utility, which is usually defined in terms of meeting a user's goals (Young et al., 2010; Rieser and Lemon, 2011). This research has shown how robust and efficient dialogue management strategies can be learned from data, but has only addressed the case of cooperative dialogue. These approaches use Reinforcement Learning with a reward function that gives positive feedback to the agent only when it meets the user's goals.

An example of the type of non-cooperative dialogue behaviour which we are generating in this work is given by agent B in the following dialogue:
A: "I will give you a sheep if you give me a wheat"
B: "No"
B: "I really need rock" [B actually needs wheat]
A: "OK... I'll give you a wheat if you give me rock"
B: "OK"

Here, A is deceived into providing the wheat that B actually needs, because A believes that B needs rock rather than wheat. Similar behaviour can be observed in trading games such as Settlers

Exp.	Learning Agent policy	Adversary policy	LA win	Adversary win
	Random	Baseline	32%	66%
a	SARSA	Baseline	49.5%	45.555%
b	SARSA + Manipulation	Baseline+Gullible	59.17%*	39.755%
1.1	SARSA+Manipulation	Basel.+ Gull.+Expos(10%).(no trade)	50.86%*	46.33%
1.2	SARSA+Manipulation	Basel.+ Gull.+Expos(5%).(no trade)	51.785%*	45.595%
2	SARSA+Manipulation	Basel.+ Gull.+Expos(10%).(win game)	49.7%	46.225%

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

of Catan (Afantenos et al., 2012).

1.1 Non-cooperative dialogue and implicature

Our trading dialogues are linguistically cooperative (based on the Cooperative Principle (Grice, 1975)) since their linguistic meaning is clear from both sides and successful information exchange occurs. Non-linguistically though they are non-cooperative, since they they aim for personal goals. Hence they violate Attardo's Perlocutionary Cooperative Principle (PCP) (Attardo, 1997).

In our non-cooperative environment, the manipulative utterances such as "I really need sheep" can imply that "I don't really need any of the other two resources", as both of the players are fully aware that three different resources exist in total and more than one is needed to win the game, so therefore they serve as scalar implicatures (Vogel et al., 2013). We have previously shown that the LA learns how to include scalar implicatures in its dialogue to successfully deceive its adversary by being cooperative on the locutionary level and non-cooperative on the perlocutionary level (Efstathiou and Lemon, 2014).

2 The Trading Game

To investigate non-cooperative dialogues in a controlled setting we created a 2-player, sequential, non-zero-sum game with imperfect information called "Taikun", between a Learning Agent (LA) and an adversary. See (Efstathiou and Lemon, 2014) for details.

Trade occurs through trading proposals that may lead to acceptance from the other player. In an agent's turn only one '1-for-1' trading proposal may occur for each resource, or nothing . Agents respond by either saying "No" or "OK" in order to reject or accept the other agent's proposal. Three manipulative actions are added to the learning agent's set of actions, of the form "I really need X" where X is a resource type. The adversary might believe such statements, resulting in modifying their probabilities of making certain trades.

2.1 Risk of exposure: Experiment 1

In this case when the Learning Agent (LA) is exposed by the adversary then the latter *does not trade* for the rest of the game. We have explored two different cases, one with a 10% chance of exposure (1.1) which gradually increases to 100% at the 10th attempt and another one (1.2) with a chance of 5%, increasing to 100% at the 20th attempt. See table 1. The results show that the LA managed to locate a successful strategy that balances the use of the manipulative actions and the normal trading actions with the risk of exposure.

2.2 Risk of exposure: Experiment 2

In this case if the LA becomes exposed by the adversary then it *loses the game*. Here we also have a 10% chance of exposure which gradually increases to 100% at the 10th attempt. See table 1. The LA learned a strategy that is similar to that of our baseline case, and it never uses manipulative actions since they are now so dangerous.

3 Conclusion & Future Work

In our previous work (Efstathiou and Lemon, 2014) we showed that a statistical dialogue agent can learn to perform non-cooperative dialogue moves in order to enhance its performance in trading negotiations. In this paper, we show that the agent can further learn how to successfully perform such moves in environments where the risk of the deception's exposure is high and the cost means either rejection of all future trades or even an instant win. Alternative methods will also be considered such as adversarial belief modelling with the application of interactive POMDPs (Partially Observable Markov Decision Processes) (Gmytrasiewicz and Doshi, 2005).

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anais Cadilhac, Cedric Degremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Verena Rieser, and Laure Vieu. 2012. Developing a corpus of strategic conversation in The Settlers of Catan. In *Proceedings of SemDial 2012*.

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

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

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

Ioannis Efstathiou and Oliver Lemon. 2014. Learning non-cooperative dialogue strategies. In *Proceedings of SIGDIAL 2014*.

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

Piotr J. Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79.

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

Elnaz Nouri and David Traum. 2014. Initiative taking in negotiation. In *Proceedings of SIGDIAL 2014*.

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

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

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

Adam Vogel, Max Bodoia, Christopher Potts, and Dan Jurafsky. 2013. Emergence of gricean maxims from multi-agent decision theory. In *Proceedings of NAACL 2013*.

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

The Disfluency, Exclamation, and Laughter in Dialogue (DUEL) Project

Jonathan Ginzburg , Ye Tian, Pascal Amsili, Claire Beyssade Barbara Hemforth, Yannick Mathieu Claire Saillard
CLILLAC-ARP (EA 3967) &
Laboratoire Linguistique Formelle (LLF) (UMR 7110)
& Laboratoire d'Excellence (LabEx)—EFL, Université Paris-Diderot, Paris, France
yonatan.ginzburg@univ-paris-diderot.fr

Julian Hough, Spyros Kousidis, David Schlangen
Faculty of Linguistics and Literary Studies,
Bielefeld University, P.O. Box 10 01 31, 33501,
Bielefeld, Germany
http://www.dsg-bielefeld.de/DUEL/

Abstract

The aim of the 3-year Disfluency, Exclamation, and Laughter in Dialogue (DUEL) project between Université Paris Diderot (Paris 7) and Bielefeld University is to model the human capacity for speaking and understanding disfluent and laughterful utterances, and to create formal models and computational systems capable of this processing. The other challenge in this enterprise is to model this interaction incrementally, that is, online as it happens word-by-word in real dialogue.

1 Introduction

Although disfluencies, exclamations and laughter occur frequently in spoken conversation, they have received little attention both within formal theories of grammar, where they are widely perceived as phenomena outside of its range, and practical dialogue modelling, where they are perceived as distractions to be filtered out. The Disfluency, Exclamation, and Laughter in Dialogue (DUEL) project, based at Université Paris-Diderot (Paris 7) and Bielefeld University aims to address this situation by an integrated empirical, theoretical, and computational research programme. The project is funded by the Agence Nationale de Recherche (ANR) and by the Deutsche ForschungGemeinschaft (DFG) within the *projets franco-allemand en sciences humaines et sociales*.

In the rest of this project description, we provide some motivation for the project and describe some of its objectives.

2 Motivation

2.1 Disfluencies in the grammar

Disfluencies are highly frequent in natural language production. They include editing terms such as *uh* and *I mean* as well as repeats—often referred to as *recycling* ('I - uh - I wouldn't', e.g. (Clark and Wasow, 1998)) and revisions. In spoken language, disfluencies are typically found in about six out of 100 words (Fox Tree, 1995) / more than 35% of all utterances (Jurafsky and Martin, 2009, p. 453).

Despite their ubiquitous nature, grammarians have, with very few exceptions, regarded disfluencies as elements not fit to populate the grammatical domain. Their very existence is a significant motivation for the competence/performance distinction (Chomsky, 1965,). (Ginzburg et al., 2014) argue that far from constituting meaningless "noise", disfluencies participate in semantic and pragmatic processes such as anaphora, conversational implicature, and discourse particles, as illustrated in (1):

(1) a. Peter was, well, he was fired. (Example from (Heeman and Allen, 1999); anaphor refers to material in reparandum.)

b. A: Because I, any, anyone, any friend, anyone, I give my number to is welcome to call me (Example from the Switchboard corpus (Godfrey et al., 1992,); implicature based on contrast between repair and reparandum: *It's not just her friends that are welcome to call her when A gives them her number*.)

c. The other one did, no, other ones did it. (Example from BNC (file KB8, line 1705); material negated by *no* originates in the reparandum.)

Beyond this, (Ginzburg et al., 2014) offer detailed argumentation for why disfluencies do belong in the grammar. In particular, they point out that disfluencies exhibit linguistic regularities across all levels of grammatical representation, cross-linguistic variation, and universals. All these

are hallmarks of processes that need representation in a grammar. Crosslinguistic variation has been documented in some detail in comparative work between morphosyntactic aspects of repair on a wide range of languages by Fox and collaborators (e.g., (Fox et al., 1996; Wouk et al., 2009; Fox et al., 2010)) and in and in phonetic analysis of hesitation markers (Candea et al., 2005,).

Understanding the range of cross-linguistic variation and the scope of universals in the area of disfluency is one of the motivations for the cross-linguistic programme of DUEL, where a parallel corpus in French, German, and Chinese will be compiled.

2.2 Laughter in the grammar

Laughter is multifunctional (Glenn, 2003). (Schegloff, 2001) illustrates the force cancelling effect of laughter (e.g. in indicating an utterance is not to be taken seriously or in enabling a socially delicate utterance to be made without causing offence.):

(2) Freda: Becaus-ah
(silence: 3.3 seconds)
Rubin: They don mind honey they're jus not gonna talk to us ever again.
Dave: =(laugh: hehem)/(ri: (h)ight)
Kathy: We don mind, we jus ne:ver gonna talk to you e:__ver (laugh) hh(h'g)
Dave: No, b't
Rubin: (laugh) hehehehehe

Laughter in its intra-sentential occurrence bears a strong relation to disfluency in enabling a speaker to express uncertainty about the force of the utterance they are making:

(3) A: [I,+I] [d,+ don't] feel comfortable about leaving my kids in a big day care center,
B: Worried that they're not going to get enough attention?
A: Yeah, and uh you know colds and things like that [laughter] (From Switchboard)

3 Objectives

3.1 Experimental Work

Interaction will be recorded in French (Paris), German (Bielefeld), and Chinese (both sites). This ensures variability both with respect to possible morphological and syntactic constraints on the placement of the phenomena of interest as well as to possible cultural differences in their discourse use. Chinese is chosen since its morphological properties lead us to expect significant variation with respect to disfluencies in polysyllabic and inflectionally rich French and German; conversely, given that the basic SVO word order of Chinese resembles French quite a bit more than German, this will also enable to control for the role of word order v. morphology.

3.2 Theoretical Work

The goal of this work area is to extend work on grounding, clarification interaction, and disfluency within the framework of KoS (Ginzburg, 2012; Ginzburg et al., 2014) l so that it can both underpin the analysis of various linguistic phenomena revolving around disfluencies, laughter, and interjections, serve as the grammar and dialogue theoretical basis for computational work, based on the parallel corpus we will have compiled. There are two main formal tasks in this area: first, develop KOS$_{incr}$, a detailed, principled incremental semantics for dialogue in KoS, using Type Theory with Records (Cooper, 2012). Second, expand KOS$_{incr}$ to KOS$^{EMA}_{incr}$, a dialogical theory whose states encode emotive appraisal (Marsella and Gratch, 2009). Each of these formal innovations will underpin detailed linguistic analysis.

3.3 Computational Work

The goal of this work area is to provide a practical, computational model of disfluency and laughter in dialogue, which captures the subtleties observed in the data and implements the main elements of the theory. The model will be implemented within an (extension of an) existing dialogue system (Buß et al., 2010), and will be evaluated for the improvements it effects in perceived naturalness of the behaviour of the system.

We aim to build a system that can be disfluent in a natural way, and is also capable of interactionally appropriate laughter when interacting with users. These are milestones for moving towards more natural spoken conversations between humans and machines.

Acknowledgements

We acknowledge support by the Lab(oratory of)Ex(cellence)–EFL (ANR/CGI) and from the Agence Nationale de Recherche (ANR) and the Deutsche ForschungGemeinschaft (DFG) within the *projets franco-allemand en sciences humaines et sociales*

References

Okko Buß, Timo Baumann, and David Schlangen. 2010. Collaborating on utterances with a spoken dialogue system using an isu-based approach to incremental dialogue management. In *Proceedings of the SIGdial 2010 Conference*, pages 233–236, Tokyo, Japan, September.

M. Candea, I. Vasilescu, M. Adda-Decker, et al. 2005. Inter-and intra-language acoustic analysis of autonomous fillers. In *Proceedings of DISS 05, Disfluency in Spontaneous Speech Workshop*, pages 47–52.

N. Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

H.H. Clark and T. Wasow. 1998. Repeating words in spontaneous speech. *Cognitive psychology*, 37(3):201–242.

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. Elsevier, Amsterdam.

B.A. Fox, M. Hayashi, and R. Jasperson. 1996. Resources and repair: A cross-linguistic study of syntax and repair. *STUDIES IN INTERACTIONAL SOCIOLINGUISTICS*, 13:185–237.

Barbara A Fox, Yael Maschler, and Susanne Uhmann. 2010. A cross-linguistic study of self-repair: Evidence from english, german, and hebrew. *Journal of Pragmatics*, 42(9):2487–2505.

Janet E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34:709–738.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):1–64.

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

Phillip J Glenn. 2003. *Laughter in interaction*. Cambridge University Press Cambridge.

John J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and devlopment. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, San Francisco, USA, March.

Peter A. Heeman and James F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing*. Prentice Hall, New Jersey, 2 edition.

S.C. Marsella and J. Gratch. 2009. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90.

E.A. Schegloff. 2001. Getting serious: Joke → serious' no'. *Journal of pragmatics*, 33(12):1947–1955.

F Wouk, B Fox, Makoto Hayashi, Steven Fincke, Liang Tao, M Sorjonen, Minna Laakso, and Wilfrido Flores-Hernandez. 2009. A cross-linguistic investigation of the site of initiation of same turn self repair. *Conversation analysis: Comparative perspectives*.

Hearer Engagement as a Variable in the Perceived Weight of a Face-Threatening Act

Nadine Glas
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI
46 rue Barrault, 75013 Paris, France
nadine.glas@telecom-paristech.fr

Catherine Pelachaud
CNRS LTCI
Télécom ParisTech
46 rue Barrault, 75013 Paris, France
catherine.pelachaud@telecom-paristech.fr

Abstract

We have performed a perceptive study in human-human interaction to verify if Brown & Levinson's formula to estimate the perceived weight of a Face-Threatening Act should be augmented with the perceived engagement level of the addressee. The outcome of this analysis will be applied to human-machine interaction, giving indications as to whether human-like virtual characters that interact with a less engaged human user should employ stronger politeness strategies than when they interact with a more engaged human user.

1 Introduction

We consider engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" (Poggi, 2007 in: Peters et al., 2005). Numerous recent studies describe how a virtual character can influence user engagement by coordinating and synchronizing its behaviour with that of its user. One of the verbal aspects that can be coordinated with the user is the degree of expressed politeness (De Jong et al., 2008). En & Lan (2012) indeed state that a successful implementation of politeness maxims is likely to improve human-agent engagement. To gain more insight into the optimal coordination of politeness, we have conducted a perceptive study to verify the existence of a link between the speaker's perceived engagement level of the hearer, and the speaker's politeness strategies.

2 Hypothesis

According to Brown & Levinson's (1987) (B&L) Politeness Theory, Wx, the numerical value that measures the weightiness, i.e. danger, of a Face-Threatening Act (FTA) x is calculated by: $W_x = D(S,H) + P(H,S) + R_x$ where $D(S,H)$ is the social distance between the speaker and the hearer, $P(H,S)$ is the power that the hearer has over the speaker, and R_x is the degree to which the FTA x is rated an imposition in that culture. The distance and power variables are intended as very general pan-cultural social dimensions. In our view, besides a very general pan-cultural distance between participants in an interaction, the level of engagement can be seen as a measure for distance as well. Considering our definition of engagement, a low level of engagement implies a temporal small value to *continue the interaction and be together with the other interaction participant(s)*. This distance may be comparable with B&L's distance variable, only this time it has a more temporal and dynamic nature. We thus formulate our hypothesis as: $W_x = D(S,H) + P(H,S) + R_x - Eng(H)$ where $Eng(H)$ is the speaker's perceived engagement level of the hearer. Related research includes André et al. (2004) who modelled an agent that takes into account the perceived emotions of the user in adapting its politeness strategy; De Jong et al. (2008) who described a model for the alignment of formality and politeness in a virtual guide; and Mayer et al. (2006) who evaluated the perception of politeness in computer based tutors.

3 Method

From B&L's theory it is apparent that a straightforward way to infer the perceived threat of an FTA is by looking at the politeness strategy that is employed to formulate it. We thus performed a perceptive study by means of a questionnaire to compare the use of politeness strategies over different conditions. Concretely, for three different FTAs (disagreement, request and suggestion), we created two conditions (written, scripted interactions) of the same scenario where two people converse, with different hearer engagement levels. We

then presented third party observers (participants of the questionnaire) with one condition of each FTA and asked them to advice the speaker (Person A) the utterance with the most appropriate politeness strategy to place the FTA, under the condition that the speaker absolutely wants to continue the conversation with the hearer (Person B). We also asked the observers to judge Person B on her level of engagement and related concepts involvement, rapport and interest.

For the context in which Person B's utterances were designed to express a minimum level of engagement we kept her utterances as brief (few and short utterances) and uninterested (emotionless) as possible. In the interactions where Person B 's utterances were designed to demonstrate a high level of engagement we added cues that have been linked to engagement in former studies and which can be expressed in written text: We made Person's B reactions longer as to extend the interaction time (Bickmore et al., 2013); we added more feedback (Gratch et al., 2006); added expressions of emotion (Peters et al., 2005.) and of liking their interaction partner (Bickmore et al., 2013); and showed interest in Person A (Peters et al., 2005).

The politeness strategies among which observers could choose were constructed according to B&L's tactics to formulate such strategies, inspired by example sentences from De Jong et al. (2008), and validated by an earlier perceptive study we performed. The validation was necessary since theoretically politeness strategies can be ranked according to their potential of minimizing the FTA's risk in the way B&L proposed, but in practice B&L's hierarchy is not always entirely respected (De Jong et al., 2008; André et al., 2004).

4 Results

200 subjects participated to our questionnaire: 68.5% female, 100% native French speakers, aged 16-75. Every participant was exposed to one version (engaged or less engaged) of each scenario (FTA). For every FTA, observers perceived the hearer's engagement, involvement, rapport and interest levels significantly higher in the engaged condition than in the less engaged condition (t-tests $p < 0.01$). Between the two conditions Mann-Whitney U tests have not shown significant differences in the distributions of recommended politeness strategies. Kendall Tau tests on the complete data set have shown significant negative correlations ($p < 0.05$), for the FTA 'request', between the rank of the chosen politeness strategy and the level of engagement ($\tau = -0.127$, Q2; $\tau = -0.111$, Q3), involvement ($\tau = -0.110$) and interest ($\tau = -0.107$). The FTA 'suggestion' holds a significant negative correlation regarding the perceived level of involvement ($\tau = -0.109$).

5 Conclusion and Discussion

In the creation of the two conditions (engaged and less engaged) we have demonstrated a successful verbal behaviour model to convey a participant's engagement level. The results do not show that the recommendation of politeness strategies differs between both conditions. The lack of such a clear overall difference confirms that politeness is a highly subjective phenomenon (Danescu-Niculescu-Mizil et al., 2013). We also compared the ranking of an observer's chosen politeness strategy for Person A with the level of engagement and related concepts he perceived in Person B. Significant negative correlations were revealed in the contexts of the negative FTAs 'request' and 'suggestion'. 'Disagreement', a threat to the addressee's positive face, does not show such correlations. A possible explanation for this is that here such a tendency interferes with a preference for alignment. Namely, a low level of engagement is expressed by features that overlap with features that indicate positive impoliteness. Some people prefer strong alignment settings and may thus be inclined to answer positive impoliteness with less caution for the addressees positive face as well (De Jong et al., 2008). The fact that the FTA 'suggestion' shows only one negative correlation may be due to the fact that the FTA can be interpreted as not really face-threatening. We conclude that in the context of a certain negative FTA, observers who choose weightier politeness strategies, tend to perceive a lower level of the addressee's engagement level, and vice versa. In these contexts, our hypothesis $W_x = D(S,H) + P(H,S) + R_x - Eng(H)$ seems confirmed, giving indications that a virtual character that wants to continue the interaction with its human user needs to speak more politely to someone who is less engaged than to someone who is very engaged in the ongoing interaction. For the future we plan to extend our study with other modalities and other aspects of engagement such as paying attention (Sidner et al., 2005) and showing empathy (Castellano et al., 2013).

Acknowledgement

We thank Brice Donval, Nesrine Fourati and Sabrina Campano for technical support and Candy Sidner for valuable discussion. We would also like to thank all the participants of our experiment. This research was funded by the French DGCIS project 'Avatar 1:1'.

References

Elisabeth André, Matthias Rehm, Wolfgang Minker and Dirk Bühler. 2004. *Endowing spoken language dialogue systems with emotional intelligence.* Affective Dialogue Systems.178–187. Springer.

Timothy W. Bickmore, Laura M. Vardoulakis Pfeifer and Daniel Schulman. 2013. *Tinker: a relational agent museum guide.* Autonomous agents and multi-agent systems.27(2):254–276. Springer.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage.* Cambridge University Press.

Ginevra Castellano, Ana Paiva, Arvid Kappas, Ruth Aylett, Helen Hastie, Wolmet Barendregt, Fernando Nabais Susan Bull. 2013. *Towards empathic virtual and robotic tutors.* Artificial Intelligence in Education.733–736 Springer.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec and Christopher Potts. 2013. *A computational approach to politeness with application to social factors.* arXiv preprint arXiv:1306.6078.

Markus De Jong, Mariët Theune and Dennis Hofs. 2008. *Politeness and alignment in dialogues with a virtual guide.* Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1.,207–214.

Looi Qin En and See Swee Lan. 2012. *Applying politeness maxims in social robotics polite dialogue.* Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on,189–190. IEEE.

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J. Van der Werf, and Louis-Philippe Morency. 2006. *Virtual rapport.* Intelligent Virtual Agents,14–27. Springer.

Richard E. Mayer, and W. Lewis Johnson, Erin Shaw and Sahiba Sandhu. 2006. *Constructing computer-based tutors that are socially sensitive: Politeness in educational software.* International Journal of Human-Computer Studies.64(1):36–42. Elsevier.

Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini and Isabella Poggi. 2005. *Engagement capabilities for ecas.* AAMAS05 workshop Creating Bonds with ECAs.

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

Candace L. Sidner, Christopher Sidner, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. *Explorations in engagement for humans and robots.* Artificial Intelligence,166(1):140–164. Elsevier.

Assessing the Impact of Local Adaptation in Child-Adult Dialogue: A Recurrence-Quantificational Approach

Robert M. Grimm
Utrecht Institute of Linguistics OTS
Utrecht University
r.m.grimm@students.uu.nl

Raquel Fernández
Institute for Logic, Language & Computation
University of Amsterdam
raquel.fernandez@uva.nl

1 Introduction

An important question in the study of dialogue is to what extent interlocutors converge on shared linguistic representations. Building on work by Dale and Spivey (2006) and Fernández and Grimm (2014), we make use of recurrence quantificational analysis (RQA) to investigate such linguistic convergence in child-caregiver dialogue. We use *convergence* as a cover term for possibly different adaptation mechanisms (e.g. priming, repetition), not all of which may be known, and without committing ourselves to the primacy of any one mechanism. However, we do assume that *convergence* is locality-dependent, since presumably the underlying mechanisms are unlikely to act on utterances that are far apart in time.

RQA (Eckmann et al., 1987) involves the construction of recurrence plots—structures which plot two data series against one another, and which allow for the extraction of further quantitative measures. We use recurrence-plot-derived measures in order to independently measure the influence of two possible constraints on the extent to which words and syntactic structures are used in both the child's and the caregiver's speech:

(1) the general use of a linguistic element in the other interlocutor's speech, and

(2) the reuse of a linguistic element in temporally close child-adult turns (i.e., *convergence*).

2 Method: Turn-based Recurrence Plots

Following Fernández and Grimm (2014), we construct turn-based recurrence plots. Given a child-caregiver dialogue, all child and adult turns are extracted; indexed by time, the child's turns are placed on the y-axis, and the adult's turns are placed on the x-axis. Every point in the resultant coordinate system then corresponds to a pair of turns. If we colour points according to the similarity of their turns, with black for maximal and white for minimal similarity, we often see a dark *diagonal line of incidence*—the set of points which compare adjacent turns. Two examples are given in Figure 1.

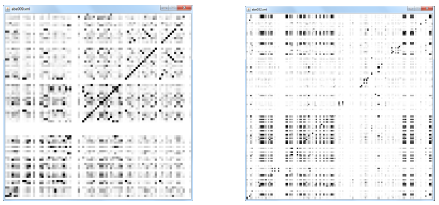


Figure 1: Recurrence plots from dialogues in the Kuczaj corpus (CHILDES database).

3 Procedure

We aim to measure the impact of factors (1) and (2) above on the frequency of linguistic elements in child and adult speech, respectively. We conduct two analyses: one focusing on the child's speech and one focusing on the adult's speech. For both analyses, we concentrate on the usage of three types of linguistic elements: content words, function words, and part-of-speech tag bigrams (POS bigrams). That way, we aim to measure a more meaning-driven usage (content words) and a more syntactically oriented usage (function words) of lexical items. POS bigrams are assumed to correspond roughly to syntactic structures and are the most syntactically oriented element type.

Our data come from three corpora in the CHILDES database (MacWhinney, 2000): Abe from the Kuczaj corpus; and Adam and Sarah from the Brown corpus. Given a linguistic element E and a dialogue, we construct a Boolean recurrence plot, where two turns are given a similarity score of 1 just in case both the child and adult turn contain E, and of 0 otherwise. We next take the sum of scores for points that correspond to turns which are at most two turns apart and which both contain E. This yields E's *raw recurrence*. Division by the frequency of E in

the dialogue yields E's *recurrence*—a frequency-independent measure of convergence. By subtracting from E's frequency the number of occurrences of E which fall within the area around the diagonal defined by $d = 2$, we similarly obtain a convergence-independent measure of frequency—E's *occurrence* (in the child's speech and in the adult's speech, respectively). In sum, given some linguistic element E, we calculate the following measures for each dialogue transcript:[1]

- *recurrence*
- *child / adult occurrence*
- *child / adult frequency*

Table 1 shows the sample sizes for each element type. Average measures over all transcripts in each corpus then form the basis of multiple linear regression models: *recurrence* and *child/adult occurrence* serve as predictors; and *adult frequency* and *child frequency* act as response variables.

corpus	cont. words	fun. words	POS big.
Abe	193 (552)	65 (172)	70 (226)
Adam	292 (553)	78 (168)	71 (195)
Sarah	254 (506)	53 (170)	50 (222)

Table 1: Sample sizes. Number of elements whose average *raw recurrence* is significantly different from the randomized condition and the total number of elements in the corpus (in parentheses).

4 Results

The regression models are summarized in Tables 2 and 3.[2] Results for the child's and adult's speech are very similar, suggesting that the two interlocutors adapt to the other's speech via the same underlying mechanisms. Regression coefficients also do not differ much across corpora, indicating that different child-caregiver dyads adapt to one another in similar ways.

Comparison of the predictor values sheds light on the impact of (1) general use in the other interlocutor's speech and (2) convergence of both interlocutors' speech in determining the frequency of a linguistic element in the child's/adult's speech. *Occurrence* takes the larger value for most element types; general use in the other's speech thus appears to have a stronger impact on the usage of content words (almost exclusively affected by occurrence) and of function words (affected by both predictors, though more strongly by occurrence). The usage of POS bigrams, lastly, is more strongly affected by *recurrence*. More syntactic elements may thus be prone to a stronger influence of convergence. We also found that the most frequent items within each element type are much more strongly affected by convergence than by general use (space constraints prevent us from elaborating on this result). Since the most frequent elements account for a disproportionately large part of the language produced, this suggest that the majority of both interlocutors' dialogue contributions may in fact be shaped through convergence.

corpus	*element type*	*adult occ.*	*recurrence*	R^2
Abe	con. words	0.79 ***	0.08	0.61
	fun. words	0.43 ***	0.48 ***	0.68
	POS big.	0.20 *	0.72 ***	0.78
Adam	con. words	0.83 ***	0.09 **	0.66
	fun. words	0.67 ***	0.34 ***	0.79
	POS big.	0.20 *	0.71 ***	0.76
Sarah	con. words	0.85 ***	0.18 ***	0.66
	fun. words	0.57 ***	0.50 ***	0.77
	POS big.	0.49 ***	0.46 ***	0.85

Table 2: Multiple linear regression models for the predictors *adult occurrence* and *recurrence*, with *child frequency* as response variable.

corpus	*element type*	*child occ.*	*recurrence*	R^2
Abe	con. words	0.80 ***	0.01	0.63
	fun. words	0.43 ***	0.54 ***	0.73
	POS big.	0.30 ***	0.65 ***	0.83
Adam	con. words	0.86 ***	0.03	0.74
	fun. words	0.73 ***	0.24 ***	0.80
	POS big.	0.23 **	0.71 ***	0.79
Sarah	con. words	0.82 ***	-0.11 ***	0.72
	fun. words	0.82 ***	-0.01	0.65
	POS big.	0.24 **	0.71 ***	0.85

Table 3: Multiple linear regression models for the predictors *child occurrence* and *recurrence*, with *adult frequency* as response variable.

5 Future Work

In future work, we aim to utilize a longitudinal design in order to track developmental changes in how specific element types are influenced by the two factors we have studied here.

[1] Importantly, we only consider E for analysis if its average *raw recurrence* differs significantly from a baseline condition where the child's turns are randomly shuffled (one-sided t-test, $p \leq 0.05$).

[2] We use the following convention to indicate significance: *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

References

Rick Dale and Micheal J. Spivey. 2006. Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3):391–430.

Jean-Pierre Eckmann, Sylvie Oliffson Kamphorst, and David Ruelle. 1987. Recurrence plots of dynamical systems. *Europhysics Letters*, 4(9):973–977.

Raquel Fernández and Robert Grimm. 2014. Quantifying categorical and conceptual convergence in child-adult dialogue. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Quebec City, Canada.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk: Volume I: Transcription format and programs, Volume II: The database*. Lawrence Erlbaum Associates.

First observations on a corpus of multi-modal trialogues

Florian Hahn, Insa Lawler & Hannes Rieser
Collaborative Research Center
"Alignment in Communication" (CRC 673)
Bielefeld University, Germany
[fhahn2|ilawler|hannes.rieser]@uni-bielefeld.de

Abstract

While there is a huge amount of work on duologues, trialogues are little investigated. We present first observations on a corpus which contains, *inter alia*, multi-modal trialogues. It turns out that we need new tools in order to do justice to the peculiarities of these forms of interactions.

1 Introduction

To communicate fluently and successfully requires humans to coordinate with each other. There are many proposals of how to analyze duologues (dialogues between two persons). Topics like turn-taking (e. g., Sacks et al. (1974)), joint project organization (e. g., Clark (1996)), and grounding (e. g., Clark and Brennan (1991), and Traum (1994)) are much discussed. But not many deal with communications beyond duologues. A notable exception is Ginzburg (2012), who, however, does not treat multi-modal utterances.

The Bielefeld Speech-and-Gesture-Alignment-corpus (short: SaGA-corpus, Lücking et al. (2013)) has been extended in order to fill this gap. The extended SaGA-corpus contains 90 duologues and 10 trialogues of participants engaged in route descriptions and/or comparisons. In the trialogues, two participants explain their routes and passed sights to a third participant, who should be able to identify both routes and the differences between them. Here, we present first observations on the essential differences between trialogues and duologues by using examples from the corpus.

2 An example for a trialogue

The two route givers (RGs) describe the beginning of the route to the so-called Follower (FO). Here, they are describing the route segment from a sculpture to another sight (the town hall). One of the RGs ("RG2") explains how to exit a roundabout (see Fig. 1).

RG2: Im Kreisel habe ich dann
In the roundabout have I then
die zweite Ausfahrt genommen
the second exit taken

FO: Also geradeaus durch
So straight ahead through,
sozusagen, oder?
so to say, right?

RG2: Genau
Exactly

RG1: Ja, das habe ich auch
Yes, that have I as well

Figure 1: Example conversation

This example is structured as follows (Fig. 2): The description by RG2 is followed by a clarification request by the FO. After that has been answered, RG1 comments by noticing that she encountered the same path at this point.

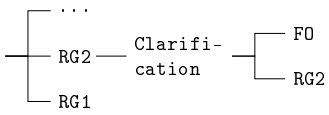


Figure 2: Structure of trialogue example

3 Essential differences between duologues and trialogues

Both duologues and trialogues require the participants to coordinate with each other to fulfill joint projects, and include a variety of communicative actions, including non-verbal actions (e.g., gestures and eye movements). However, there are crucial differences of trialogues to duologues.

3.1 Differences concerning joint projects

For our analyses of the conversations, we follow Clark's conception of a joint project (JP). "A joint project is a joint action projected by one of its participants and taken up by others" (Clark (1996): 191), whereas a joint action is an action carried out by more than one person (e. g., dancing a waltz). The overall joint-project of the trialogues in our corpus is the comparison of two routes and sights described by RG1 and RG2 to a FO. A big JP as this one is realized by several smaller JPs. Each JP is characterized by two actions: an action by one of the participants (e. g., a question) and the reaction/response of the others (e. g., an answer).

The main differences between duologues and trialogues concerning JPs lie in the responses. Firstly, the common *binary* adjacency pair organization is not applicable to most JPs. An example is a question requiring two answers by different participants. One also needs group acceptance in order to initiate and complete joint projects of the group. It would not suffice if only one or two participants agree. In our trialogues, especially the comparisons of route segments are acknowledged by all of the participants before the route description continues. In our example, both FO (after the clarification request) and RG1 acknowledge the description by RG2. This observation can be substantiated with numerous corpus examples.

Secondly, the scope of acknowledgements can differ. While in duologues it can be assumed that the scope of an acknowledgement extends over (parts of) the last contribution, the acknowledgements in trialogues can also extend over more than one contribution. Take one example: One of the RGs tells the FO "The fact by which you can recognize it [the townhall] easily is simply that there are two little trees next to the door". Next, the other RG claims "Ah, right. They were [there] as well", by which she presumably means that there were also two little trees on her ride through the town. Then, the FO says "Ah, trees", whereby she acknowledges *both* utterances.

Thirdly, the differences in responses are crucial for grounding. If you get acceptance in a duologe the resulting mutual belief of the agents can be based on individual beliefs in the manner of epistemic logics. However, in trialogues you can have different groupings of agents and then you need a notion of group belief which cannot be reduced to individual beliefs (see Rieser (2014) for a systematic overview on individual and group beliefs).

3.2 Differences concerning turn-taking

The current addressee in common duologues is the non-talking participant. There is usually no need for an explicit addressing. In trialogues one always has to explicitly address the addressee of one's contribution if it is not addressed to both participants in order to avoid confusion. If one does not use proper names to do that, one can achieve it by using eye contact or gesture, or by employing context information. In our example, the addressee of the question is RG2 because the clarification request is clearly related to his description.

This difference in addressing also has an influence on turn-taking regularities. The projection of the end of a turn and turn transition relevance points (Sacks et al., 1974) presumably works in the same way as in duologues. But the taking of a turn is organized differently, because in absence of explicit addressing there are two potential turn takers. In our trialogues, one influence on turn-taking is the kind of role of the respective participant. The FO is expected to ask questions about route segments and the sights (beyond clarification requests). Thus, it is easier for her to win the turn-taking competition. The turn-taking also depends on the overall organization of the joint project realization. Depending on the kind of structure used, there are certain expectations about who's turn is next. For instance, in consecutive Route-Sight(RS)-comparison (Fig. 3 in appendix) it is expected that RG2 takes the floor after RG1 has finished his/her description (including clarification requests). Similar rules can also be given for other kinds of RS-comparisons (Figures 4 & 5 in app.). Such an expectation does not apply to the comparison-phases. Since all are required to compare the descriptions, there is no one preferred.

4 Conclusion

Our first observations strongly suggest that there are peculiar features of trialogues which need to be modelled by extending the common tools for analyzing duologues. In our future research, we will provide fine-grained analyses of trialogues in the extended SaGA-corpus to gain a deeper understanding of the phenomenon. We also want to stress the role that gestures play in the organization of trialogues, and aim to build here on our work on discourse gestures (Hahn & Rieser, 2011).

Appendix

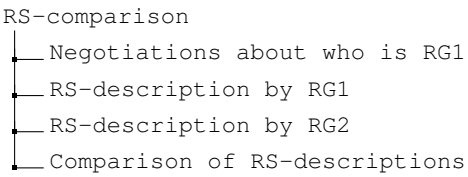


Figure 3: Consecutive RS-comparison

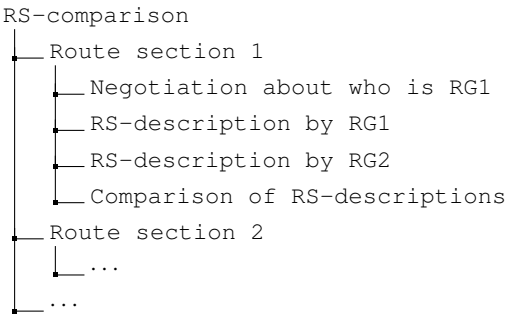


Figure 4: Consecutive RS-comparison step by step

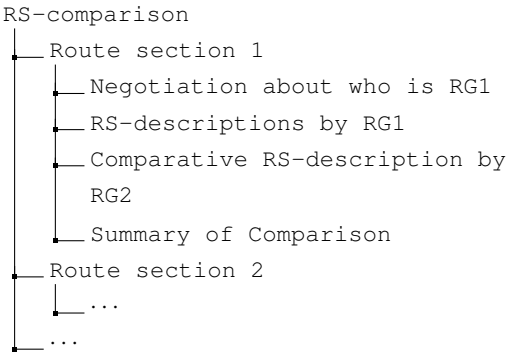


Figure 5: Immediate RS-comparison

Acknowledgements

The work was supported by the German Research Foundation in the CRC 673 "Alignment in Communication".

References

Herbert Clark. 1996. *Using Language*. Oxford University Press.

Herbert Clark, Susan Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition,* 13, 127–149.

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

Florian Hahn, Hannes Rieser. 2011. Gestures Supporting Dialogue Structure and Interaction in the Bielefeld Speech and Gesture Alignment Corpus (SaGA) *Proceedings of SemDial 2011*: 182–183.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2): 5–18.

Hannes Rieser. 2014. *Implementing Grounding into Dialogue Theory*. Bielefeld University, Ms.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50: 696–735.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation.*. Ph.D. thesis, Rochester, NY, USA.

Towards Automatic Understanding of 'Virtual Pointing' in Interaction

Ting Han
Bielefeld University

Spyros Kousidis
Bielefeld University

David Schlangen
Bielefeld University

firstname.lastname@uni-bielefeld.de

1 Introduction

When trying to convey, from memory, the placement of objects relative to each other, one can use descriptions such as "the one is about two centimeters to the left of the other, and roughly one centimeter higher", or one can just place ones hands in a representation of this configuration and say something like "one is *here* and the other one is *here*".

The type of gesture used in these latter displays has been called "abstract dexis" (McNeill et al., 1993) or "virtual pointing" (Kibrik, 2011), and it has been observed that these gestures have the remarkable effect of *creating* extralinguistic spatial referents for objects that are mentioned in the discourse, but are not in fact currently present. These referents can later in discourse be used to re-refer to the same entity; in our example, this could be done via "and *this* one [accompanied by pointing gesture] is".

Lascarides and Stone (2009) make the interesting proposal that such gestures do indeed call attention to a real location in shared space (which they denote with variables such as $\vec{p}$), but carry their semantic load via a mapping (v) into the conveyed location ($v(\vec{p})$) in the described situation, where the identity of the mapping is contextually determined. Configurations of locations indicated via such gestures (e.g. a $\vec{p_1}$ and a $\vec{p_2}$) then achieve their iconic value as a depiction of a configuration between the locations they are mapped into ($v(\vec{p_1}), v(\vec{p_2})$).

We were interested in how stable over time and how precise in their iconicity such mappings are in actual instances of use, with a view at how automatic understanding of such speech/gesture ensembles could be realized. We elicited and recorded multimodal spatial scene descriptions, and measured precision by fitting a mapping between virtual referent locations and true object locations. We then used this mapping to retrieve from the set of all scenes the one that was being described. Using our matching method, we find that the gestures carry a good amount of spatial information for 45 out of 53 episodes. In current work, we are attempting to make this retrieval process incremental, and combine it with an understanding of the utterance that the gestures accompany.

2 The Corpus

In order to elicit pointing gestures in a virtual space, we designed a simple description task in which participants were shown an image on a computer screen for a brief time (10 seconds) and then were asked to describe it.

The images showed a configuration of four objects, and an arrow indicating a movement of one of the objects; this movement was also to be described. An example of such an image is shown in Figure 1. The objects were always simple geometric shapes, and at most two different colors were used. The scenes were designed in such a way that if gestures were used to indicate locations, this would have to be done successively (as there were more objects than hands available to the subjects), and that for at the very least one object, namely the one that is to undergo the motion, there would be a need for a repeated reference.

For all participants, the same series of 50 images was used that each is different from the others, but a time limit of 20 minutes was set for the whole experiment, and several participants did not complete the full set.

In total, we recorded 311.63 minutes of video (by a HD camera) and motion capture data (by Leap motion sensor[1]. Since we are interested in shapes in 2D, we only analyzed the data in x-y plane for all 3D data collected by Leap sensor.),

[1] www.leapmotion.com

of which 179.51 minutes contain speech. 14 participants took part in the experiment, each of them finished 29 scene descriptions on average (SD = 9.60). The analyses below were performed on 53 episodes (with 4 original references) from 8 dialogues, as not all data is annotated yet.

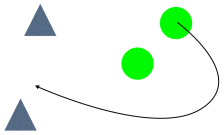


Figure 1: One of the scenes used in the experiments

3 Shape Matching and Scene Retrieval

Shape Matching The four objects in a scene form a shape with 4 vertexes, which can be represented as a matrix:

$$S = \begin{Bmatrix} x1 & y1 \\ \vdots & \vdots \\ x4 & y4 \end{Bmatrix} \quad (1)$$

in which rows correspond to object positions.

After getting the detected virtual pointing shape, we want to know how close it is to the original shape (S_o). However, due to different personal gesture space and pointing behaviors, the two shapes are not identical. We performed a shape matching method[2] to transform the detected shape to a target shape (S_t) which is most close to the original shape by shifting, rotating and scaling the detected shape, an example is shown in Fig 2.

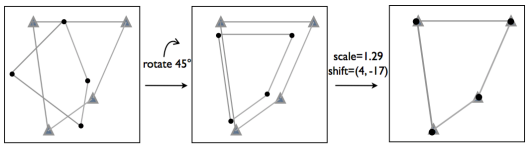


Figure 2: Shape matching

First of all, a randomly initialized transform parameter vector p is generalized:

$$p = [\theta, t_x, t_y, s] \quad (2)$$

where θ is the rotating angle; t_x and t_y stand for the shift value on x and y axis; s is the scaling parameter. For each row in matrix S we do rotation, shift and scaling with following equation:

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

By minimizing the cost function:

$$E = min \parallel S_t - S_o \parallel \quad (4)$$

we get an optimized p which can transform the detected shape to the target shape S_t. We evaluate how close the detected shape is to the original shape with matching error, which is the distance between the target shape and the original shape.

Scene Retrieval We matched each detected virtual pointing shape with each of the 50 scenes that were prepared and ranked matching errors in ascending sequence. With good iconicity in the gestures, the matching error between virtual pointing shape and the original scene should have a low rank value, and the shape should pick out the scene that was actually described in the given episode from the set of all scenes that have been described.

4 Results and Discussion

For 21 of all episodes (39.62%), the gestured shape shows the smallest matching error among all candidates. In 30 episodes (56.60%), the gestured shape has error rank two. Consequently, the 2-best accuracy of using gesture shape information to retrieve the described scene is an impressive 96.22%. The remaining 2 episodes had a matching error above rank 2. A random selection baseline on this task would give an 1-best accuracy of 1.88%.

To fully evaluate these results, they would need to be weighted with a measure of similarity between the scenes that were to be described (because distinguishing between similar scenes based on spatial information is more difficult than between wildly different ones). But even in this form, the results already indicate that the gestures carry fairly accurate information about one aspect of the described scene. We take this as a starting point for our current work of combining this gestural information, in an incremental fashion, with information from the utterances that it accompanies (Kennington et al., 2013). The next step then will be to model recreation of the scenes from scratch, rather than selection from a set of candidates.

[2] http://glowingpython.blogspot.com/2013/06/shape-matching-experiments.html

References

Casey Kennington, Spyridon Kousidis, and David Schlangen. 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. Proceedings of SIGdial 2013.

Andrej A. Kibrik. 2011. *Reference in discourse*. Oxford University Press, Oxford, UK.

Alex Lascarides and Matthew Stone. 2009. A Formal Semantic Analysis of Gesture. *Journal of Semantics*, 26(4):393–449.

David McNeill, Justine Cassell, and Elena T. Levy. 1993. Abstract deixis. *Semiotica*, 95(1-2):5–20.

Two Alternative Frameworks for Deploying Spoken Dialogue Systems to Mobile Platforms for Evaluation "In the Wild"

Helen Hastie, Marie-Aude Aufaure*, Panos Alexopoulos, Hugues Bouchard, Heriberto Cuayáhuitl, Nina Dethlefs, Milica Gašić, Almudena González Guimeráns, James Henderson, Oliver Lemon, Xingkun Liu, Peter Mika, Tim Potter, Verena Rieser, Pirros Tsiakoulis, Yves Vanrompay, Boris Villazon-Terrazas, Majid Yazdani, Steve Young and Yanchao Yu

email: h.hastie@hw.ac.uk. See `http://parlance-project.eu` for full list of affiliations

Abstract

We demonstrate two alternative frameworks for testing and evaluating spoken dialogue systems on mobile devices for use "in the wild". We firstly present a spoken dialogue system that uses third party ASR (Automatic Speech Recognition) and TTS (Text-To-Speech) components and then present an alternative using audio compression to allow for entire systems with home-grown ASR/TTS to be plugged in directly. Some advantages and drawbacks of both are discussed.

1 Introduction

This abstract describes the EC FP7 PARLANCE project whose goal is to perform interactive search through speech in multiple languages. With the advent of evaluations "in the wild", emphasis is being put on converting research prototypes into mobile applications that can be used for evaluation and data collection by real users downloading the app from the market place. This is the motivation behind the work demonstrated here. We present a modular framework whereby research components from the PARLANCE project (Hastie et al., 2013) can be plugged in, tested and evaluated in a mobile environment. The domain is interactive search for restaurants in San Francisco, USA. All required restaurant information is obtained through a Yahoo search API which returns entities based on their longitude and latitude within San Francisco for 5 main areas, 3 price categories and 52 cuisine types containing approximately 1,600 individual restaurants.

2 Two System Architectures

The first framework adopts a client-server approach as illustrated in Figure 1 for the PARLANCE system in Mandarin (Hastie et al., 2014). This system uses third party Google ASR and TTS, where the recognised utterance is sent to the Stanford Segmenter[1] server and the segmented utterance is then sent to the Spoken Language Understanding (SLU), Interaction Manager (IM) (Thomson and Young, 2010), Natural Language Generation (NLG) (Dethlefs et al., 2013) and TTS components in sequence. For details of all the PARLANCE components please see (Hastie et al., 2013) and the project website[2].

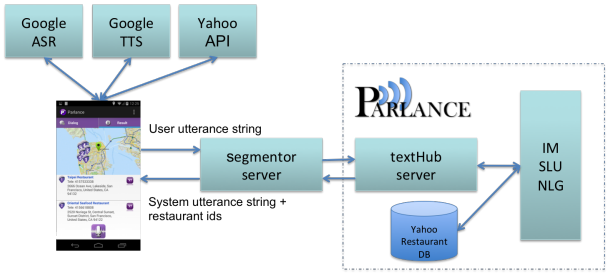


Figure 1: Architecture 1: the PARLANCE Mandarin mobile application system architecture using third party ASR/TTS.

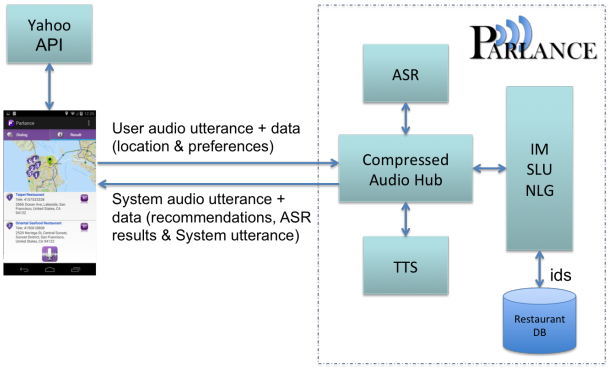


Figure 2: Architecture 2: the PARLANCE English mobile application system architecture using audio compression.

*Authors are in alphabetical order

[1] http://nlp.stanford.edu/projects/chinese-nlp.shtml

[2] http://parlance-project.eu

However, we were also interested in integrating and evaluating all the PARLANCE capabilities, such as user barge-in and incrementality (Hastie et al., 2013) and did not want to rely on third party software. Therefore, we developed an alternative architecture for the English version using a SIP client-server communication. However, this proved sensitive to bandwidth variations and some carriers and Internet service providers block it. The final version avoids this problem by transferring highly compressed audio and data using internet connectivity as illustrated in Figure 2.

Similar dialogue system frameworks also make use of audio compression for network-based ASR (Pieraccini et al., 2002) and TTS (Kruijff-Korbayová et al., 2012). They also transfer audio files (but without compression) for network-based ASR and use either a server TTS (Gruenstein et al., 2008) or a client TTS (Fuchs et al., 2012). Others train language understanding components from crowdsourcing based on speech input and output components running on a server (Liu et al., 2013).

2.1 Discussion of Architectures

Advantages of the first architecture include rapid development and easy portability to new domains. This is due to off-the-shelf components being used which save effort in development and testing. This is true for dialogue systems in multiple languages, where home-grown ASR/TTS do not exist. An advantage of the second architecture is that home-grown and domain-specific ASR and TTS components can often lead to better performance than off-the-shelf components (Dušek et al., 2014; Tsiakoulis et al., 2014). However, reasonable response times per turn should be taken into account (between 100 and 500 milliseconds) (Strömbergsson et al., 2013). Another advantage of the second architecture is that is allows incremental processing for input analysis and output planning. This has been shown to lead to more natural interactions that human users prefer over their non-incremental counterparts (Skantze and Schlangen, 2009).

2.2 Multimodal Functionality

In addition to spoken dialogue, the mobile app features substantial multi-modal interaction functionality. It displays the set of results during the conversation with the system and allows refinement and inspection of the results while talking. Hyper-local features include being able to sort results by distance from the user and also organised by neighbourhoods or nearby Points-of-Interest (POIs) (Bouchard and Mika, 2013). This last feature is particularly appealing in a tourism scenario where the user may not be aware of neighbourhoods in the city, but might remember the location of major sights. Screenshots of the English mobile app (Architecture 2) are shown in Figures 3 and 4.

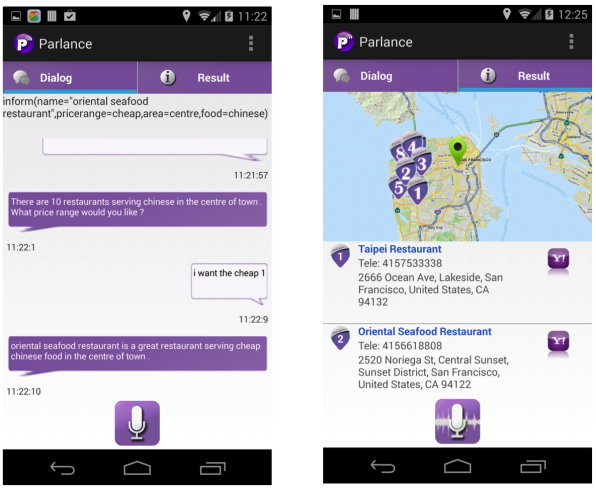


Figure 3: Screenshot of a dialogue and the list of recommended restaurants also shown on a map.

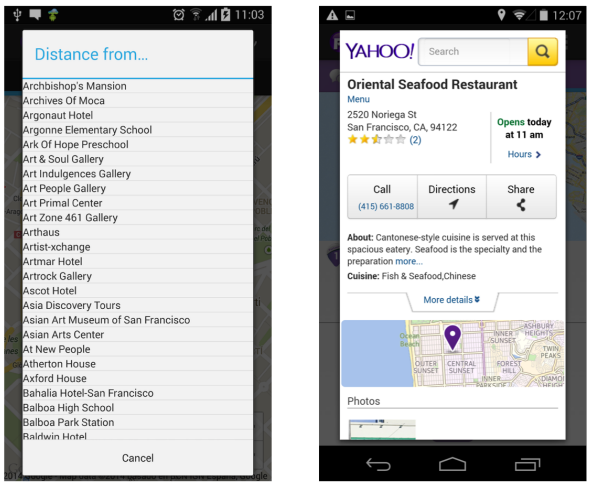


Figure 4: Screenshot of recommended restaurants and ordering by distance from points of interest.

3 Future Work

Future work involves developing a feedback mechanism for evaluation purposes that does not put undue effort on the user and put them off using the application. In addition, this framework could be extended to leverage social information of the user when displaying items of interest.

Acknowledgements

The research leading to this work was funded by the EC FP7 programme FP7/2011-14 under grant agreement no. 287615 (PARLANCE).

References

H. Bouchard and P. Mika. 2013. Interactive hyperlocal search API. Technical report, Yahoo Iberia, August.

N. Dethlefs, H. Hastie, H. Cuayáhuitl, and O. Lemon. 2013. Conditional Random Fields for Responsive Surface Realisation Using Global Features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

O. Dušek, O. Pltek, L. Žilka, and F. Jurčíček. 2014. Alex: Bootstrapping a spoken dialogue system for a new domain by real users. In *Proceedings of SIGDIAL*, Philadelphia, PA, U.S.A.

M. Fuchs, N. Tsourakis, and M. Rayner. 2012. A scalable architecture for web deployment of spoken dialogue systems. In *Proceedings of LREC*.

A. Gruenstein, I. McGraw, and I. Badr. 2008. The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of ICMI*.

H. Hastie, M.A. Aufaure, P. Alexopoulos, H. Cuayáhuitl, N. Dethlefs, M. Gasic, J. Henderson, O. Lemon, X. Liu, P. Mika, N. Ben Mustapha, V. Rieser, B. Thomson, P. Tsiakoulis, Y. Vanrompay, B. Villazon-Terrazas, and S. Young. 2013. Demonstration of the PARLANCE system: a data-driven incremental, spoken dialogue system for interactive search. In *Proceedings of SIGDIAL*, Metz, France, August.

H. Hastie, M.A. Aufaure, P. Alexopoulos, H. Bouchard, C. Breslin, H. Cuayáhuitl, N. Dethlefs, M. Gašić, J. Henderson, O. Lemon, X. Liu, P. Mika, N. Ben Mustapha, T. Potter, V. Rieser, B. Thomson, P. Tsiakoulis, Y. Vanrompay, B. Villazon-Terrazas, M. Yazdani, S. Young, and Y. Yu. 2014. The PARLANCE mobile application for interactive search in English and Mandarin. In *Proceedings of SIGDIAL*, Philadelphia, PA, U.S.A.

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 *Proceedings of WOCCI*.

J. Liu, P. Pasupat, S. Cyphers, and J. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *Proceedings of ICASSP*.

R. Pieraccini, B. Carpenter, E. Woudenberg, S. Caskey, S. Springer, J. Bloom, and M. Phillips. 2002. Multimodal spoken dialog with wireless devices. In *Proceedings of ISCA Tutorial and Research Workshop - Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany.

G. Skantze and D. Schlangen. 2009. Incremental Dialogue Processing in a Micro-Domain. In *Proceedings of EACL*, Athens, Greece.

S. Strömbergsson, A. Hjalmarsson, J. Edlund, and D. House. 2013. Timing responses to questions in dialogue. In *Proceedings of INTERSPEECH*.

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

P. Tsiakoulis, C. Breslin, M. Gasic, M. Henderson, D. Kim, M. Szummer, B. Thomson, and S. Young. 2014. Dialogue context sensitive hmm-based speech synthesis. In *Proceedings of ICASSP*.

Disentangling utterances and recovering coherent multi party distinct conversations

Gibson O. Ikoro
School of Electronic Engineering and Computer Science, Queen Mary University of London
g.o.ikoro@qmul.ac.uk

Raul Mondragon
School of Electronic Engineering and Computer Science, Queen Mary University of London
r.j.mondragon@qmul.ac.uk

Graham White
School of Electronic Engineering and Computer Science, Queen Mary University of London
graham.white@qmul.ac.uk

Abstract

Automatic separation of interposed sequence of utterances into distinct conversations is an essential prerequisite for any kind of higher-level dialogue analysis. Unlike most models that involve highly computational intensive methods such as clustering techniques, our proposed approach uses a simple and efficient sequential thread detection method which is less computational intensive. It uses the waiting time (time gap between the current speaker and the next speaker), similarity between utterances, turn-taking and participant-based features.

```
1
2  40:29 A→(B):grins I think it's the proxy servers
3               called Kevin and Perry that need kicking!
4  40:55 B→(A):what happened last night..the whole
5               fecking lot of it got or needed a kicking!
6  41:13 C→(D,E,F,G,H): lsaysl cH kissing bandit...l
7  41:45 H→(I,J):Kissing bandits are  predators and
8                should not be tolerated
9  41:46 A→(B): it was a Janet router that went again,
10              second tie in a week that one has died for
11 42:08 C→(D,E,F,G,H):lsaysl cYou're just jealous he
12              took your job
13 42:16 A→(B):grins...janet is the nae of the network
14              that the universities and schools are on.
15              A router is soething that forwards on
16              inforation to the correct coputer, so
17              when you send your essage one TCZ, lo
18 42:21 H→(I,J): And I haven't gotten any action since
19
```

Figure 1: Sample of conversation from our corpus.

1 Motivation:Disentanglement problem

Chat rooms are where people can meet each other to chat on the internet (Davies, 2010). Chat room logs are not a single continuous conversation of two or a group of people at a time rather each time widow is a sequence of frequently broken utterances (Elsner and Charniak, 2008). A typical conversation, therefore, does not form an adjacent segment of the chat-room transcript, but sequences of frequently broken utterances due to interposed utterances from other conversations. For example, consider the time window (40:29 - 42:21) mm:ss in Fig 1 the utterances in line 2, 4, 9 and 13 show an ongoing conversation which is being interposed by the utterances in line 6, 7, and 11. However, a typical chat room log that consists of millions of utterances and conversations which interpose each other will be difficult to separate into distinct conversations using traditional methods.

Another challenge in disentangling chat log is schism, a process where some participants create a new conversation which is different from the already existing one. This often occurs when two or more users change their attention to themselves and away from whoever held the floor (the current speaker) in the parent conversation (Elsner and Charniak, 2008). Disentanglement or thread detection is a task that extracts the different interposed utterances in a chat log and separates them into distinct conversations. If we want our statistical methods to be useful for conversational analysis, we have to disentangle the logs. It is only when we have disentangled the logs, that we can apply other methods to find out about structures like question-answer pairs (Purver, 2011).

However, any form of automated semantic analysis is tedious and likely to be unsuccessful on account of the extremely unstructured lexicon used (Camtepe et al., 2005). Hence there is need for a model that combines both the pragmatic information and statistical approach.

2 Related work

Thread disentanglement is most commonly studied using clustering methods (Uthus and Aha, 2013; Elsner and Charniak, 2008). In a recent study, (Elsner and Charniak, 2011) employed coherence models to investigate chat disentangle-

ment. They validated their models using recorded telephone conversations for thread disentanglement. In their method they used tabu search method to search for a solution to this problem; this involves conducting two sets of experiments with each chat corpus. The first is to disentangle single messages and the second disentangles the entire chat log.

Another interesting method on chat disentanglement is described in (Elsner and Schudy, 2009) and (Elsner and Charniak, 2010). Their work utilized correlation clustering for thread detection. The method involves searching for group of clusters that maximizes the degree of similarities between pairs within a cluster and maximizes the degree of dissimilarity among pairs across clusters. The two models employ maximum-entropy classifier to determine if two messages belong to the same conversation. Elsner and Schudy employ two methods: greedy method and local search method for the NP-hard problem of searching the best solution for correlation clustering while Elsner and Charniak, employed voting schema for correlation clustering (Uthus and Aha, 2013).

In another recent work, Mayfield et al. (Mayfield et al., 2012) utilized a two-pass method for thread detection. In the first pass, the method labels sentences using a negotiation framework. After the labelling process, a single-pass clustering algorithm is used to detect sequences.

Our approach is unique in the sense that it does not involve any conventional clustering method or other highly computational intensive techniques which may lead to depreciation in the accuracy of results. Before discussing our proposed techniques, we will introduce the dataset.

3 Description of the dataset

Walford is a text-based online social community that was set up more than a decade ago (Healey et al., 2008). It has roughly 2446 regular users. It is a corpus that contains 24040 hours (26/11/2001 - 24/08/2004) of chat. For each communication, the following data is recorded: the time, the originator, the originators location, the recipient(s) and their location.

In Walford, the participants can construct a friend-list. Walford has a tool that permits users to send direct message to all the members in their friend list who are online at the same time (Healey et al., 2008). The ability to reach everyone in one's friend list simultaneously helps Walford users to perform group chat.

4 Methodology

The proposed approach for the ongoing project uses a simple and efficient method for chat disentanglement. The algorithm involves three-pass process. In the first pass, the algorithm predicts the occurrence of schism and use turn-taking allocation rule and timing to extract the users who are involved in the schism. In the second pass, the algorithm separates the individual utterances to form different datasets using the waiting time (time gap) and turn-taking allocation rule. In the third pass, The algorithm recovers a complete distinct conversation thread from the utterances by looking at the participants-based features and the content similarity between the utterances.

4.1 Schism detection

There are two ways in which new conversations can start, one is through a schism and the other is through a conversation initiating statement. According to (Uthus and Aha, 2013) "Schism occurs when a conversation splits into two conversations the new conversation is formed due to certain participants branching off from a specific message and refocusing their attention upon each other". This implies that the users who are involved in schism were once an audience of the current speaker in the main conversation before the schism occurred and secondly, the two conversations seems to occur at the same time. With these features we can predict when and where schism starts.

4.2 Waiting time

We considered the waiting time or time gap in a chat room communication as the time difference between successive messages. The waiting time is calculated using approach in (Mihaljev et al., 2011) as

$$dt = t_{i+1} - t_i$$

,

where t_i is the time at i and t_{i+1} is the time at i+1.

For example in Fig 1, the waiting time or time gap between $A \rightarrow B$ and $B \rightarrow A$ in line 2 and 3 respectively is 29 seconds $(40:55 - 40:29)$ and the waiting time or time gap between $B \rightarrow A$ and

$C \rightarrow (D, E, F, G, H)$ in line 3 and 4 respectively is 44 seconds $(40 : 29 - 41 : 13)$.

Since we know the temporal distribution of waiting time in a given conversation, we use this data to estimate the likelihood of a particular utterance belonging to a given conversation.

4.3 Content and participant based features

The content based features will involve comparing the number of word-similarity between two utterances. For example the number of words shared between utterance X and utterance Y suggests that the two utterances may belong to the same conversation (Joty et al., 2013). The participant-based features is described as follows:

- Pairs or group of utterances X and Y may be closely connected in the discourse and are likely to be directly related if those participating in utterance X are the same people participating in utterance Y and the time between them falls within the extracted waiting time distribution.

- Pairs or group of utterances X and Y may be widely separated in the discourse and are unlikely to be directly related if those participating in utterance X is totally different from those people participating in utterance Y.

5 Summary

We have proposed a simple and efficient approach for chat disentanglement. It will avoid using methods that are highly computational intensive, instead it uses simple data characteristics such as utterance similarities, response waiting time, turn-taking and the participant-based feature. With this approach, we hope to achieve results that will be nearer-human performance on an annotated corpus.

References

Seyit Ahmet Camtepe, Mark Goldberg, Mukkai Krishnamoorty, and Malik Magdon-ismail. 2005. Detecting conversing groups of chatters: a model, algorithms, and tests. In *In Proceedings of the IADIS International Conference on Applied Computing*, pages 89–96.

Faith Davies. 2010. The history of chat rooms.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio, June. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Comput. Linguist.*, 36(3):389–409.

Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1179–1189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Micha Elsner and Warren Schudy. 2009. Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

PatrickG.T. Healey, Graham White, Arash Eshghi, AhmadJ. Reeves, and Ann Light. 2008. Communication spaces. *Computer Supported Cooperative Work (CSCW)*, 17(2-3):169–193.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2013. Topic segmentation and labeling in asynchronous conversations. *J. Artif. Int. Res.*, 47(1):521–573, May.

Elijah Mayfield, David Adamson, and Carolyn Penstein Ros. 2012. Hierarchical conversation structure prediction in multi-party chat. In *In Proceedings of SIGDIAL Meeting on Discourse and Dialogue*.

T. Mihaljev, L. de Arcangelis, and H. J. Herrmann. 2011. Interarrival times of message propagation on directed networks. , 84(2):026112, August.

Matthew Purver. 2011. Topic segmentation. In G. Tur and R. de Mori, editors, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317. Wiley.

David C. Uthus and David W. Aha. 2013. Multiparticipant chat analysis: A survey. *Artificial Intelligence*, 199200(0):106 – 121.

Large-scale Analysis of the Flight Booking Spoken Dialog System in a Commercial Travel Information Mobile App

Zengtao Jiao
Baidu Inc.
Beijing, China

Zhuoran Wang
Heriot-Watt University
Edinburgh, UK

Guanchun Wang
Baidu Inc.
Beijing, China

Hao Tian
Baidu Inc.
Beijing, China

Hua Wu
Baidu Inc.
Beijing, China

Haifeng Wang
Baidu Inc.
Beijing, China

Abstract

In this paper, we analyze around three hundred thousand real user dialogs collected from a publicly deployed flight booking spoken dialog system (SDS), to investigate the correlations between the task completion rate and user locations and daily time periods, as well as the correspondences between user responses and system requests. The findings can serve as guidelines to design more granular strategies for SDS in this domain.

1 Introduction

Due to the recent advances of mobile technology and the prevalence of smart devices in the latest decade, commercialized speech interfaces, particularly spoken dialog systems (SDS), are gaining increasing popularity. Successful examples include Apple's Siri, Google Now and Microsoft Cortana, to name just a few. The broad deployment of such applications enables more advanced analyses of SDS based on a vast amount of data generated by real users in real-world scenarios. Previous studies of this kind can be found in (Williams, 2011; Williams, 2012).

This paper studies several interesting phenomena observed in a large-scale data set of real user dialogs collected from a flight booking SDS developed by Baidu and integrated in a travel information mobile app widely used in China. The SDS here is a rule-based system following the RavenClaw architecture (Bohus and Rudnicky, 2009), where the dialog manager only takes top SLU hypotheses into account when making decisions.

2 Data Analysis

The data analyzed in this work consist of around 300K dialog sessions and more than 600K turns collected from our SDS during the first half of 2014. Basic statistics show that the task completion rate[1] for these dialogs is 77%. Based on such data, two factors that may affect the task completion rate are investigated in detail, including user's departure/destination locations and time periods of a day when the dialogs occur. In addition, we also investigate the correspondences between user responses and systems requests, which reflects user habits and the properness of each system action.

Departures and Destinations We cluster user's departure and destination cities according to the provinces they belong to, and plot the province-wise departure and destination task completion rates in Figure 1. Firstly and very interestingly, it can be found that the three most popular tourist provinces, Hainan, Yunnan and Tibet, demonstrate exactly opposite effects to the task completion rate when they occur as the departure and the destination locations. To explain this phenomenon, one can imagine that a user using the flight booking system at a tourist place would tend to have a clear goal in mind (e.g. searching for a flight back home), whilst in many cases the users searching for flights to a tourist place may just want to browse flights and compare prices without any specific plan in mind, especially the travel date (which results in 57.8% of the task failures according to our statistics). It suggests that a better dialog policy should consider user intentions adaptively when knowing the above prior knowledge, rather than treat all the destinations uniformly. Secondly, for some provinces, such as Hebei and Anhui, the task completion rate is relatively low, regardless of them being the departure or the destination locations. This may be because the ASR is less ro-

[1] There are 3 required slots (departure, destination and flight date) in the SDS, which must be filled before the system can execute a database search. The task completion rate defined here stands for the percentage of dialogs where all the three required slots are filled. There are 14 optional slots not considered when computing the task completion rate here.

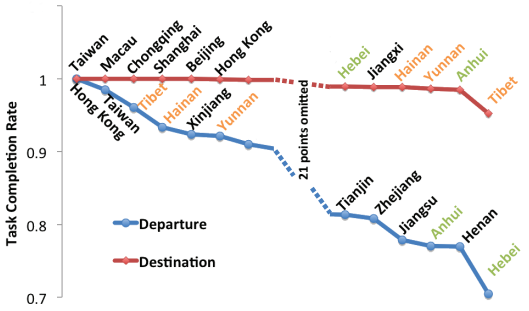


Figure 1: Task completion rate with respect to departure and destination provinces.

bust to the accents or dialects in those provinces. Note that, as our SDS will initialize the departure place according to user's GPS location if such information is available, most of those failed dialogs can still have their departure slots filled by default. Therefore, in the above figure, the task completion rates for departure provinces tend to be lower than those for destination provinces.

Daily Time Periods We also analyze the task completion rate of our system with respect to different daily time periods, as shown in Figure 2. It can be found that highest task completion rates occur during the midnight till early morning, and it decrease significantly in the evening, where the valley points are observed around 6pm and 8pm. It can be understood that people using the system in "abnormal" time periods (such as midnight to early morning) may have a strong requirement and motivation to have a journey booked. But in the evening (such as 8pm), one would expect that many users may just play with the app for entertaining purposes. A more attractive interaction strategy could be identifying those entertaining intentions and addressing them in a less formal manner. Environmental noise will be another factor affecting the task completion rate (e.g. the peak traffic hours 6pm∼7pm). A noise-level prior would improve the robustness of the SDS, particularly if a statistical system (Young et al., 2013) is employed in the future.

User Responses vs. System Requests Based on SLU-parsable utterances only, we investigate the correspondences between system requests and user responses, for which the results are illustrated in Figure 3. The statistics here aim to reflect user habits and to further examine the properness of the design of our system actions. It suggests that the users rarely say departure place and date in one

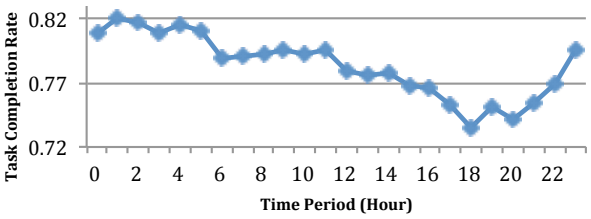


Figure 2: Task completion rate with respect to daily time periods.

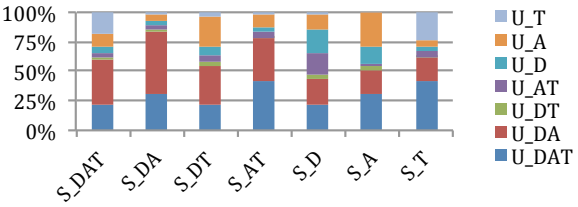


Figure 3: Correspondences between system requests (S) and user responses (U) for the three slots, departure (D), destination (A) and date (T).

utterance, therefore, the system may not sensibly benefit from asking for all such information simultaneously. Similar but slightly better correspondence is found for destination in conjunction with date as well.

3 Conclusion

This work discusses potential improvements to the granularity of SDS based on large-scale real user data analyses, for which the practical solutions will be the focus of our future work.

Acknowledgements

The research in this paper is supported by the 973 Program No. 2014CB340505. The second author is supported in part by a SICSA PECE grant.

References

D. Bohus and A. I. Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. *Comp. Speech Lang.*, 23(3):332–361.

J. D. Williams. 2011. An empirical evaluation of a statistical dialog system in public use. In *SIGDIAL*.

J. D. Williams. 2012. Challenges and opportunities for state tracking in statistical spoken dialog systems: Results from two public deployments. *IEEE J. Selected Topics Sig. Proc.*, 6(8):959–970.

S. Young, M. Gasic, B. Thomson, and J. Williams. 2013. POMDP-based statistical spoken dialogue systems: a review. *Proceedings of the IEEE*, PP(99):1–20.

A Multi-issue Negotiation Model of Trust Formation through Concern Alignment in Conversations

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

'Concern Alignment in Conversations' project aims to investigate the relationship between rational agreement seeking and affective trust management through conversations. The project conducts both empirical analyses of real-life conversation data that involve both agreement and trust, e.g., various types of consultation conversations, including medical domain dialogues, and computational modeling of the processes connecting agreement seeking and trust management taking place behind those conversational exchanges.

Our guiding idea in the project is the notion of *'concern alignment'*, that aims to schematically capture conversational processes from the perspective of consensus-building and trust formation (Katagiri et al., 2011; Katagiri et al., 2012).

2 Trust through conversation

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 (Clark, 1996). 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. We identify 'trust' as a type of mental states that enables us to form, even lacking sufficient support, presumptive expectations on other agents' choice of actions, and to choose our own actions based on those presumptions.

3 Concern alignment

We conceptualize dialogue consensus decision-making processes as consisting of two functional parts, concern alignment and joint plan construction, as shown in Figure 1. When a group of people engage in a conversation to find a joint course of actions among themselves on certain objectives (*issues*), they start by expressing what they deem relevant on the properties and criteria on the actions to be settled on (*concerns*). When they find that sufficient level of alignment of their concerns is attained, they proceed to propose and negotiate on concrete choice of actions (*proposals*) to form a joint action plan.

We have been iteratively developing a set of dialogue acts (Allen and Core, 1997; Bunt, 2006) for concern alignment through annotating real-life consultation conversations and refining the dialogue act set.

4 Analysis of concern alignment

Figure 2 shows an annotation example of a part of a medical obesity counseling dialogue session. The analysis captures the process of concern alignment in which the nurse A tries to identify all the possible concerns related to the smoking behavior the patient B by both her own concern introduc-

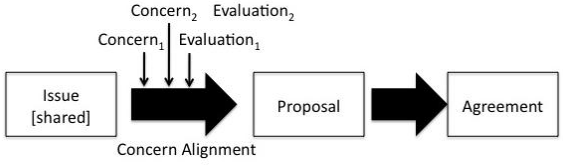


Figure 1: A schematic diagram of the concern alignment process in consensus-building dialogues.

A-B:	C-solicit:	(what makes you want to smoke)		
A-B:	C-introduce:	(when offered)	⇒	C-eval/negative:(do not smoke even when offered)
A-B:	C-introduce:	(with somebody smoking)	⇒	C-eval/positive: (sometimes I will)
A-B:	C-introduce:	(feel impatient)	⇒	C-eval/positive: (I do smoke when I feel impatient)
A-B:	C-introduce:	(with tea or coffee)		
B-A:	C-introduce:	(when drinking)		
		⇓		
A-B:	P-solicit:	(when with somebody smoking?)		
B-A:	P-introduce:	(will leave the place)		
A-B:	P-solicit:	(when you feel impatient?)		
B-A:	P-introduce:	(can manage if I have something in my mouth)		
A-B:	P-elaborate:	(how about e-cigar?)		
B-A:	P-reject:	(tried but failed)		
A-B:	P-introduce:	(how about stop-smoking pipe?)		
B-A:	P-accept:	(I've wanted to try)		
A-B:	P-solicit:	(when drinking?)		
B-A:	P-introduce:	(the same [stop-smoking pipe])		

Figure 2: An example analysis of sequential organization of concern/proposal exchanges.

Participant A					Participant B			
issue	weight	estimate	align		align	estimate	issue	weight
$Concern_1$	u_1^A						$Concern_1$	u_1^B
$Concern_2$	u_2^A						$Concern_2$	u_2^B
$\vdots$	$\vdots$	$\hat{u_i^B}$	w_i^A	$\Leftrightarrow$	w_i^B	$\hat{u_i^A}$	$\vdots$	$\vdots$
$Concern_i$	u_i^A						$Concern_i$	u_i^B
$\vdots$	$\vdots$						$\vdots$	$\vdots$
$Concern_n$	u_n^A						$Concern_n$	u_n^B

Figure 3: Concern alignment as multi-issue negotiation

tion and elicitation from the patient. A then tries to draw out proposals from B to refrain from smoking for each and all of the concerns raised.

5 Trust through multi-issue negotiation

Multi-issue negotiation: We have been exploring several models to capture and describe the conversational processes of concern alignment in computational terms, including the one based on the idea of multi-issue negotiation (Traum et al., 2008). Assuming that conversational participants A, B have their own utility u_i^A, u_i^B for each of the issues $Concern_i$ (Figure 3). The process of concern exchange in concern alignment for a participant A can be modeled by the process of estimating the utility structure on multiple issues $\hat{u_i^B}$ of your interlocutor B through the exchange of information on their own utility structures u_i^A and u_i^B.

Joint utility maximization: In the phase of negotiation on joint action proposals, participants have to weigh the utility structure of their interlocutors by the weight w_i against their own utilities. Participants then propose a joint action which maximizes the combined utilities. Alignment in this phase can be captured as the adjustment of alignment weight w_i.

Trust as parameters for joint utility: Under this concern alignment as multi-issue negotiation picture, trust can be conceived to correspond to parameters for joint utility computations. Once the process of concern alignment succeeds in obtaining a mutually satisfactory consensus, parameters such as interlocutor utility structure estimate $\hat{u_i}$ and alignment weight w_i can be utilized in later consensus-building negotiations. This accumulation of parameter values through a successful concern alignment history constitutes one's trust in others in the sense that it provides the basis in forming reasonable expectations on the interlocutor behavior choices in succeeding interaction sessions.

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

James F. Allen and Mark G. Core. 1997. DAMSL: Dialog act markup in several layers (Draft 2.1). Technical report, Multiparty Discourse Group, Discourse Resource Initiative.

Harry Bunt. 2006. Dimensions in dialogue act annotation. In *the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

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.

David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Intelligent Virtual Agents: Lecture Notes in Computer Science*, pages 117–130.

Multimodal Dialogue Systems with InproTK$_S$ and Venice

Casey Kennington
CITEC, Dialogue Systems Group, Bielefeld University
ckennington[1]

Spyros Kousidis
Dialogue Systems Group
Bielefeld University
spyros.kousidis[2]

David Schlangen
Dialogue Systems Group
Bielefeld University
david.schlangen[2]

[1]@cit-ec.uni-bielefeld.de
[2]@uni-bielefeld.de

Abstract

We present extensions of the incremental processing toolkit INPROTK which, together with our networking adaptors (*Venice*), make it possible to plug in sensors and to achieve situated, real-time, multimodal dialogue. We also describe a new module which enables the use in INPROTK of the Google Web Speech API, which offers speech recognition with a very large vocabulary and a wide choice of languages. We illustrate the use of these extensions with a real-time multimodal reference resolution demo, which we make freely available, together with the toolkit itself.

1 Introduction

In face-to-face conversation, interlocutors normally do more than just listen to speakers: they also observe what speakers do while they speak, for example how they move and where they look. Sensors that can make such observations are becoming ever cheaper. Integrating (i.e., fusing) the data they provide into the understanding process, however, is still a technical challenge (Atrey et al., 2010; Dumas et al., 2009; Waibel et al., 1996). We illustrate how our InproTK$_S$ suite of tools (Kennington et al., 2014) can make this process easier, by demonstrating how to plug together a little demo tool that combines instantiations for motion capture (via *Leap Motion*,[1]), eye tracking (*eyetribe*[2]) and speech (Google Web Speech).[3]

Furthermore, truly multimodal systems are more feasible today then they were 5 or 10 years ago, due to the proliferation of affordable sensors for common requirements in multimodal processing, such as motion capture, face tracking and eye tracking, among others. However, each sensor is typically constrained to specific platforms and programming language, albeit mostly the most common ones, a fact that hinders integration of such sensors into existing spoken dialogue systems. InproTK$_S$ and our Venice tools are a step towards streamlining this process.

In this paper, we will briefly describe INPROTK and the extensions in InproTK$_S$. We will then describe Venice and give a use case, which we have packaged into a real-time working demo.

2 The IU model, INPROTK

As described in (Baumann and Schlangen, 2012), INPROTK realizes the *IU*-model of incremental processing (Schlangen and Skantze, 2011; Schlangen and Skantze, 2009), where incremental systems consist of a network of processing *modules*. A typical module takes input from its *left buffer*, performs some kind of processing on that data, and places the processed result onto its *right buffer*. The data are packaged as the payload of *incremental units* (IUs) which are passed between modules.

3 Extensions of InproTK$_S$

InproTK$_S$ provides three new methods of getting information into and out of INPROTK:

- *XML-RPC*: *remote procedure call* protocol which uses XML to encode its calls, and HTTP as a transport mechanism.[4]
- *Robotics Service Bus*: (RSB), a message-passing middelware (Wienke and Wrede, 2011).[5]

[1] https://www.leapmotion.com/
[2] https://theeyetribe.com/
[3] We also have instantiations for *Microsoft Kinect* and *Seeingmachines FaceLAB*, www.seeingmachines.com/product/facelab/
[4] http://xmlrpc.scripting.com/spec.html
[5] https://code.cor-lab.de/projects/rsb

- *InstantReality*: a virtual reality framework, used for monitoring and recording data in real-time.[6]
- *Google Web Speech* has also been implemented as a module, in a similar manner to (Henderson, 2014).[7]

The first three methods have implementations of *Listeners* which can receive information on their respective protocols and package that information into IUs used by InproTK$_S$. Each method also has a corresponding *Informer* which can take information from an IU and send it via its protocol. A general example can be found in Figure 1, where information from a motion sensor is sent into InproTK$_S$ (via any of the three methods), which packages the information as an IU and sends it to the NLU module; later processed information is sent to an informer which then sends it along its protocol to an external logger.

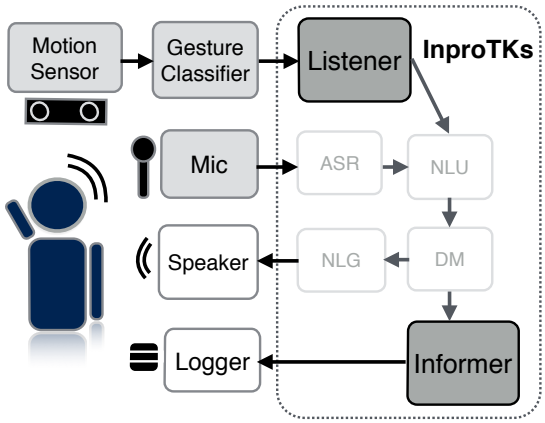


Figure 1: Example architecture using new modules: motion is captured and processed externally and class labels are sent to a listener, which adds them to the IU network. Arrows denote connections from right buffers to left buffers. Information from the DM is sent via an Informer to an external logger. External gray modules denote input, white modules denote output.

4 Bridging components together: Venice

Our venice components allows integration of any sensor software quickly and easily into InproTK$_S$ by using either of the RSB and InstantIO protocols described above as a network bus. *Venice.ipc* is a platform independent service that accepts data on a socket and pushes it to the network bus. Thus, the sensor SDK can be in any language/major OS and still be quickly integrated. *Venice.hub* is a central component that allows IO to/from any of the two protocols and disk, and is thus used for synchronous logging of all the data on the network, as well as replaying and simulating. Any number of components (sources and/or targets) can be added/removed from such a network at runtime. The Listener/Informer components of InproTK$_S$ communicate directly to this network for multimodal data I/O. Components can reside on the same computer or on dedicated workstations in a LAN.

5 Use case: The Multimodal Reference Resolution Demo

Using InproTK$_S$ we have developed a spoken dialogue system that performs online reference resolution in the Pentomino domain using three modalities: speech, gaze and deixis. We use the Leap sensor for motion capture and eyetribe for eye tracking. Both sensors are used by modifying one of their SDK examples with minimal effort, in order to send data to the venice.ipc service running on the machine. The latter sends the data using InstantIO to InproTK$_S$. The application that uses the toolkit has two InstantIO Listeners (one for each modality) and a Listener for the ASR (Google Web Speech). These are effortlessly connected to the main module (that performs the reference resolution) by means of an XML configuration file.

The main module itself performs the fusion by distributing probabilities to different candidate referents based on the input from each modality independently. If data from different modalities point to different candidates, a flat probability distribution occurs, with no candidate significantly more likely to be the referent. If more than one modalities point to the same candidate, then its probability overcomes a threshold and the reference is resolved. The confidence distribution is output by InproTK$_S$ via an Informer module back to the network and is displayed in real-time by a separate component (a Virtual Reality browser).

[6] http://www.instantreality.org/

[7] https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html

References

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. *Multimodal fusion for multimedia analysis: a survey*, volume 16. April.

Timo Baumann and David Schlangen. 2012. The InproTK 2012 Release. In *NAACL*.

Bruno Dumas, Denis Lalanne, and Sharon Oviatt. 2009. Multimodal Interfaces : A Survey of Principles , Models and Frameworks. In *Human Machine Interaction*, pages 1–25.

Matthew Henderson. 2014. The webdialog Framework for Spoken Dialog in the Browser. Technical report, Cambridge Engineering Department.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. InproTKs: A Toolkit for Incremental Situated Processing. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 84–88, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2009. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of the 10th EACL*, number April, pages 710–718, Athens, Greece. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialoge & Discourse*, 2(1):83–111.

Alex Waibel, Minh Tue Vo, Paul Duchnowski, and Stefan Manke. 1996. Multimodal interfaces. *Artificial Intelligence Review*, 10(3-4):299–319.

Johannes Wienke and Sebastian Wrede. 2011. A middleware for collaborative research in experimental robotics. In *System Integration (SII), 2011 IEEE/SICE International Symposium on*, pages 1183–1190.

Producing Verbal Descriptions for Haptic Line-Graph Explorations

Matthias Kerzel, Özge Alaçam, Christopher Habel
Department of Informatics
University of Hamburg
Hamburg/Germany
{kerzel,alacam,habel}@informatik.uni-hamburg.de

Cengiz Acartürk
Cognitive Science
Middle East Technical University
Ankara/Turkey
acarturk@metu.edu.tr

1 Verbally Assisted Haptic Graphs

Combining haptic graphs and verbal information in a multimodal human-computer interaction scenario is a promising means to make statistical graphs accessible to blind people. For example, users can explore haptic graphs by hand-controlled movements using a stylus of a force-feedback device gathering information about the graphs geometrical properties (Acartürk, Alaçam & Habel, 2014). In the present paper we look on haptic graph exploration as a collaborative activity of two agents, a (visually impaired) explorer (E) of a haptic graph and an observing assistant (A) providing verbal assistance (Habel, Alaçam & Acartürk, 2013; Alaçam, Acartürk & Habel, 2014; see Fig. 1). In particular we focus on one technical aspect in building a common ground between human explorers and computational assistants. (Sect. 2 & 3).

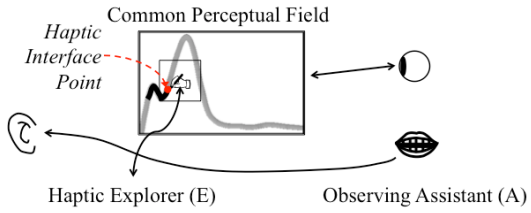


Figure 1. Assisted haptic graph exploration

In our empirical studies (Alaçam, Acartürk & Habel, 2014), the agents share a common field of perception, namely the haptic graph and its exploration, but their perception and comprehension processes differ substantially. For example, while E explores the segment of the haptic graph on a horizontal virtual plane that is the highlighted, black in Fig. 1, A perceives visually the global shape of the graph and E's exploration movement on a vertical computer screen.

The success of verbally assisted haptic graph comprehension depends on the alignment of the interlocutor's internal models, especially on building implicit common ground (Garrod & Pickering, 2004), as well as, on producing adequate utterances. The recognition of exploration events by the verbal assistant system is one of the crucial processes that ground alignment and make the communication between E and A efficient and effective.

2 Recognition of Exploration Events

For giving verbal assistance, A has to observe E's exploration movements, with other words, beyond considering the current location of the exploration movement, i.e., the haptic interface point (see Fig. 1), A has to analyze the ongoing exploration event. And in the long run, A has to consider the history of exploration events as well as the history of produced utterances.

In the following we give a short overview of how exploration movements are recognized in our OBSERVINGASSISTANT prototype:

- The haptic-graph knowledge base contains spatial information about the 3D haptic model of the specific line graphs, information about the their geometry (Kerzel & Habel, 2013) and also line-graph specific information, like positions and properties of graph landmarks like extrema and slopes.
- Haptic exploration-event recognition is performed by a rule based system (RBS) in a two-step process:
  - *Recognition of basic events*, i.e., of events that are perceivable, momentary changes in movement behavior or position with regard to segments or graph-landmarks in the virtual haptic environment. Detected basic events are inserted into the knowledge base.
  - *Rule-based complex event processing:* From perceived basic events, a rule based system, interprets basic events and constructs complex events. This system is realized in

DROOLS using a rule language based on first order predicate logic with an added temporal calculus (see, Kerzel, 2013).

- Recognized exploration events and also updates to ongoing exploration events can be used to trigger system reaction. In the current stage of development using canned text assisting speech is realized using the text-to-speech platform MARY (Schröder & Trouvain, 2003).

3 The ObservingAssistant at Work

In the following we exemplify the processes of event recognition with a short sequence of exploration movements. Figure 2 depicts, firstly, a segmented and annotated graph as it is represented in the haptic-graph knowledge base, and secondly, a segment from an exploration movement of a user (depicted by red lines); the user can not perceive the representational features depicted in Figure 2.

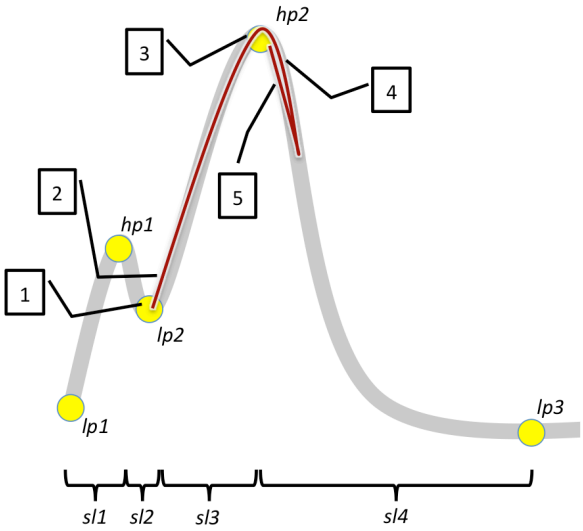


Figure 2. Analyzed exploration trajectory (red) and depictions of some content in the haptic-graph knowledge base. The numbered boxes correspond to the steps discussed in this section.
Representational features: Yellow dots ≈ local / global extremum (*lp* – minima, *hp* – maxima); horizontal curly brackets ≈ projection of *slopes* (*sl*) to the x-axis.

Step 1: The exploration starts with the haptic interface point being positioned in the local extremum *lp2*, where two slopes, namely *sl2* and *sl3* meet. By beginning to move to the right (upwards) a basic event of being in this slope is detected. The RBS recognizes that this slope is currently explored and inserts the resulting exploration event into the knowledge base. Additionally, this event can trigger canned text frames that lead for example to. "*You are exploring a steep slope in the first quarter of the graph.*"

Step 2: The user's stylus movement goes on along the slope *sl3* in the right upward direction. A basic movement event including the movement direction is detected and therefore accessible by the RBS. Considering the geometrical properties of the currently explored slope, which are stored in the knowledge base, the next relevant graph landmark that the user is approaching is identified by RBS as *hp2*. This triggers further verbal information in a look-ahead style: E.g. "*You approach the second maximum of the graph. It is the global maximum.*"

Step 3: The user reaches the maximum *hp2*. A corresponding basic event is detected. This event is also subsumed under the ongoing exploration of the slope *sl3*. The RBS reasons that the ongoing exploration event is completed since both endpoints of the slope were visited. This leads to uttering: "*You fully explored the upward slope.*"

Also an ongoing exploration event of the high point is created. Thus a further, domain dependent specification of the explored graph-landmark can be verbalized: *"You reached the global maximum of 94."*

Step 4: The user's exploration goes on beyond the maximum and enters the next slope, *sl4*, which is recognized by the RBS: An extended exploration event regarding this slope is created and together with the movement information (the user is still moving from left to right) assistance is triggered: "*You are now descending a very steep slope.*"

Step 5: The user stops and moves back to the high point *hp2*. When the user reaches the maximum, an exploration of *hp2* is recognized again. Due to the RBS'—currently even rudimentary—bookkeeping of exploration events and utterances, the assistance "*This is the global maximum again."* is given.

4 Conclusion

Haptic line-graph comprehension can be enhanced by verbal assistance. The rule-based ObservingAssistant reported above analyzes the users' exploration movements and triggers reactively canned text, which is realized by the MARY text-to-speech system. The next version of the ObservingAssistant will be extended on the basis of empirical studies on human-human assistance (such as, Acartürk, Alaçam, & Habel, 2014; Alaçam, Acartürk & Habel, 2014).

References

Acartürk, C., Alaçam, Ö., & Habel, C. (2014). Developing a Verbal Assistance System for Line Graph Comprehension. In A. Marcus (Ed.): *Design, User Experience and Usability* (DUXU/HCII 2014), Part II, LNCS 8518. (pp. 373–382). Berlin: Springer-Verlag.

Alaçam, Ö., Acartürk, C., & Habel, C. (2014). Referring Expressions in Discourse about Haptic Line Graphs. To be published in Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue. SemDial 2014 – DialWatt. Verena Rieser & Phillipe Muller (eds.)

Garrod, S., & Pickering, M. J. (2004). Why is Conversation so Easy? *Trends in Cognitive Sciences, 8*, 8-11.

Habel, C., Alaçam, Ö., & Acartürk, C. (2013). Verbally Assisted Comprehension of Haptic Line-Graphs: Referring Expressions in a Collaborative Activity. In *Proceedings of the CogSci 2013 Workshop on the Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference (PRE-CogSci 2013)*.

Kerzel, M. (2013). Rule patterns for event recognition during exploration of hapic virtual environment line-based graphics. Technical report, Department of Informatics, University of Hamburg, Germany. http://www2.informatik.unihamburg.de/wsv/pub/Kerzel-COSITsupplemental-2013.pdf

Kerzel, M. & Habel, C. (2013). Event recognition during the exploration of line-based graphics in virtual haptic environments. In T. Tenbrink, J. Stell, A. Galton & Z. Wood (eds.) *Spatial Information Theory, 11th International Conference, COSIT 2013*. (pp. 109–128). Berlin: Springer-Verlag.

Schröder, M. & Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology, 6, 365-377.*

Effects of Speech Cursor on Visual Distraction in In-vehicle Interaction: Preliminary Results

Staffan Larsson
Simon Dobnik
Sebastian Berlin
University of Gothenburg
Box 200, SE-405 30 Gothenburg, Sweden
sl@ling.gu.se, simon.dobnik@gu.se, sebastian.berlin@gu.se

Abstract

This paper presents preliminary results on visual distraction tests concerning various multimodality solutions for in-vehicle dialogue systems in the SIMSI project. In particular, the Speech Cursor concept is evaluated in comparison to other solutions and is found to decrease visual distraction, especially for tasks involving list browsing.

1 Background

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.

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 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 Speech Cursor technology[1] (Larsson et al., 2011). By reading out the item currently in focus, 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 functionality.

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).

2 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 is 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 four variants were tested:

1. GUI only (haptic only in, graphics only out)
2. GUI with speech cursor (haptics only in, graphichs and speech out)
3. Multimodal with speech cursor (haptics and speech in, graphichs and speech out)
4. Speech-only with speech cursor (haptics and speech in, speech only out)

[1] The combination of Speech Cursor and spoken dialogue interaction is Patent Pending.

For each condition, there are two difficulty levels: (1) easy and (2) 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.

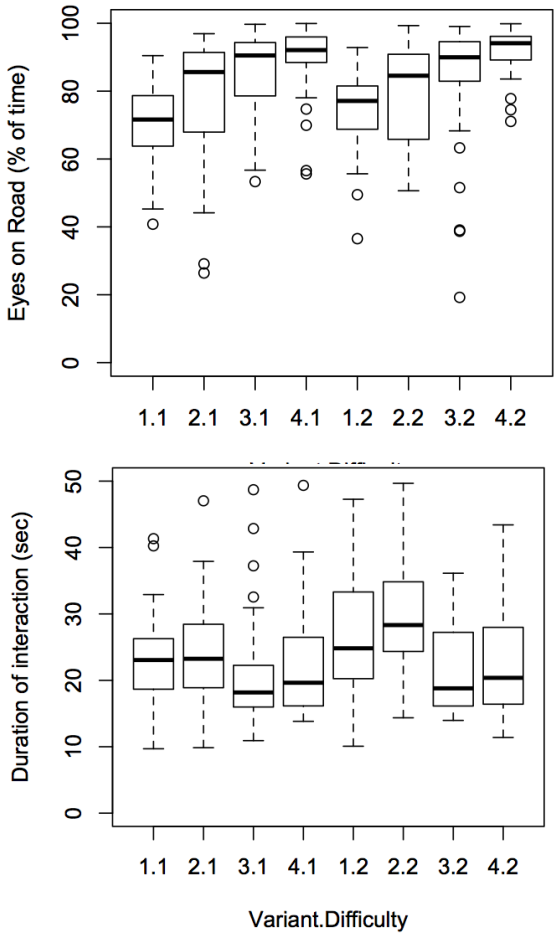
This experimental setup, which we informally refer to as the "annoying cars" setup, differs from existing experimental setups such as the ConTRe task (Engonopoulos et al., 2008). In the latter, the driver tries to match two vertical lines representing the vehicle's position and the target (reference) 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).

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 driver should then carry out this instruction as efficiently and completely as possible.

3 Results

This section presents results in the form of box plots[2]. The first box plot shows the % of time spent looking at the road in the different multimodality variants (the first number, as explained above), and difficulty levels (the second number, where 1=easy and 2=difficult). The second box plot shows the duration of interactions.

Even without spoken input, the Speech Cursor solution (variant 2) does better than GUI-only system (variant 1) w.r.t. visual distraction. Spoken input further (variant 3 and 4) reduces visual distraction, and reduces interaction time. The same trend was observed for both difficulty levels. The effect of modality condition on % Eyes on road has been tested with ANCOVA (with participant ID as co-variable) and was found to be significant at level $p < 0.001$.



4 Discussion

From these preliminary observations, we can tentatively conclude that in tasks which require browsing, the Speech Cursor will significantly decrease visual distraction while browsing compared to a GUI only solution. This is true regardless of whether the system has spoken dialogue capabilities or not, at least insofar as spoken dialogue is not used for browsing[3].

The effect of this on overall visual distraction in in-vehicle interaction will depend on the amount of browsing carried out in an interaction, which in part will depend on the nature of the domain. For example, it's more common to browse for restaurants than to browse for who to call.

As the data is skewed, the normality assumption for statistical testing cannot be maintained and therefore we intend in future work to use statistical tests that are not dependent on this assumption, such as for example Generalised Linear Mixed Models (GLMMs).

[2]For an explanation of box plots, see e.g. `http://en.wikipedia.org/wiki/Box_plot`.

[3]For example, Apple's voice-controlled CarPlay system requires the driver to look at the screen when browsing lists.

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.

Common Ground and Joint Utterance Production: Evidence from the Word Chain Task

Jarosław Lelonkiewicz
University of Edinburgh,
School of Philosophy, Psychology and Language Sciences,
7 George Square, Edinburgh, EH8 9JZ
J.R.Lelonkiewicz@sms.ed.ac.uk

Chiara Gambi
Saarland University
Psycholinguistics Group
Building C 7.4, Campus Saarbrücken,
Saarbrücken, 66123
gambi@coli.uni-saarland.de

Abstract

Instances in which one interlocutor continues an utterance initiated by another are not infrequent in conversation. Yet, the factors influencing their occurrence are not fully understood. Employing a novel experimental paradigm, this study investigated whether it is easier to jointly produce an utterance that refers to something in the common ground. We found that participants interacting via a text chat-tool had a higher typing speed for non-ambiguous than ambiguous words. This result shows that lack of shared knowledge negatively affects joint language production.

1 Introduction

Cross-person completions are considered evidence for the collaborative and incremental nature of dialogue and have attracted considerable attention amongst researchers over the last two decades (Clark, 1996; Hayashi, 1999; Helasvuo, 2004; Poesio & Rieser, 2010; Howes, Purver, Healey, Mills, & Gregoromichelaki, 2011). Despite their importance for understanding the mechanisms governing conversation, there are very few experimental studies looking into the factors facilitating joint utterance production. One valuable exception is a study by Howes, Healey, Purver & Eshghi (2012). Among other things, these authors found that participants were more likely to continue an artificially truncated turn when it was about the current conversation topic than when it introduced a new topic.

Interestingly, it is possible that the reported preference is related to interlocutors finding it easier to predict one another's utterances when these utterances are about something in common ground (see Pickering & Garrod, 2013). The aim of our study was to directly test the hypothesis that joint production would proceed more smoothly when interlocutors are talking about something in common ground.

2 Design

In our study, participants were asked to jointly produce definitions of English words. We manipulated experimentally whether the meaning of the word was in common ground, or not, by varying whether the to-be-defined word was unambiguous (i.e., had only one meaning) or ambiguous (i.e., had at least two meanings). Specifically, we used 20 ambiguous ($M_{CELEXfrequency}$ = 1107 p.m.; M_{length} = 1.4 syllables) and 20 non-ambiguous words ($M_{CELEXfrequency}$ = 1211 p.m.; M_{length} = 1.4 syllables; p's > .2), that were closely matched for frequency and length in number of syllables. The ambiguous words were balanced (dominant meaning frequency ≤ .65 and ≥ .41). We hypothesised that joint production would proceed more smoothly if participants were able to assume shared meanings with their partner (as should be the case with non-ambiguous words), because this would constrain their predictions about what will be uttered next.

3 Methods

Eighteen pairs of participants were tested. Participants were seated in separate booths and interacted via a text-based chat environment. The task was implemented using DiET chat-tool (http://cogsci.eecs.qmul.ac.uk/diet/; Mills & Healey, submitted), allowing to log key presses and typing times with great precision. The participants were presented with one word at a time and had to jointly construct a definition for

each word. Instructions emphasised speed, but also that the definition had to provide sufficient information to allow a third party to guess the word. As in the American TV game Chain Reaction, the participants could contribute only one word per turn, and had to continuously switch turns with their partner (see 1, produced as a definition of BAT). Although natural joint production lacks such a constraint, it similarly requires incremental interpretation and tight yoking of comprehension and production processes.

(1) **A**: Baseball - **B**: tool - **A**: that - **B**: is - **A**: used - **B**: to - **A**: hit - **B**: the - **A**: ball

As a control, 26 participants provided definitions for the same words in a solo version of our task. Similarly to those working together, solo participants could type only one word per turn, but were working entirely on their own.

4 Results

We measured the total time spent typing and the number of words produced per definition, and computed typing speed as number of words per second. Data were analysed using linear mixed effects models (Baar, Levy, Scheepers, & Tily, 2013), as implemented in the lme4 package (Bates, Maechler, & Dai, 2008). Significance of the fixed effects was assessed by means of likelihood ratio tests.

Typing speed was higher for non-ambiguous than ambiguous words when participants were interacting with another ($M_{non\text{-}amb}$ = .50, M_{amb} = .46), but not in the solo task ($M_{non\text{-}amb}$ = .71, M_{amb} = .71; Ambiguity X Task interaction: $\chi 2$ (1) = 4.21, $p < .05$; maximal random effects structure). This suggests that lack of shared knowledge negatively affects the joint performance at the task.

5 Discussion

We showed that jointly producing an utterance is more difficult when common ground cannot be assumed but needs to be established. Note that our dependent variable (typing speed) should primarily index ease of language production. Therefore our study provides further insight into mechanisms governing dialogue, and adds to the existing evidence for the role of common ground in comprehension (e.g., Brown-Schmidt, 2009).

Additional analyses should investigate whether typing speed is affected predominantly at the beginning of definitions for both ambiguous and non-ambiguous items. This would confirm that the observed difference in typing speed reflects the cost of establishing common ground. It would also provide further support for the hypothesis that the information about what is shared between speakers influences the prediction of the upcoming turn of the interlocutor.

Acknowledgements

We would like to thank Gregory Mills for help with the chat-tool.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Maechler, M., & Dai, B. (2008). The lme4 package. [Computer software manual]. Retrieved from http://cran.r-roject.org/web/packages/lme4/lme4.pdf

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialogue. *Journal of memory and language*, *61*(2), 171-190.

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

Hayashi, M. (1999). Where Grammar and Interaction Meet: A Study of Co-Participant Completion in Japanese Conversation. *Human Studies*, *22*(2), 475–499.

Helasvuo, M.-L. (2004). Shared syntax: the grammar of co-constructions. *Journal of Pragmatics, 36*(8), 1315–1336.

Howes, C., Healey, P. G., Purver, M., & Eshghi, A. (2012). Finishing each other's… Responding to incomplete contributions in dialogue. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 479-85).

Howes, C., Purver, M., Healey, P. G., Mills, G., & Gregoromichelaki, E. (2011). On incrementality in dialogue: Evidence from compound contributions. *Dialogue & Discourse, 2*(1), 279-311.

Lerner, G. H. (2004). Collaborative turn sequences. In *Conversation analysis: Studies from the first generation*, 225–256. John Benjamins.

Mills, G. J., & Healey, P. G. T. (submitted). A dialogue experimentation toolkit. Retrieved from http://cogsci.eecs.qmul.ac.uk/diet/

Pickering, M. J. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329-347.

Poesio, M. & Rieser, H. (2010). Completions, coordination, and alignment in dialogue. *Dialogue and Discourse, 1*(1), 1-89.

Language-bound dialogic elements in Computer-Mediated and Face-to-Face Communication

Barbara Lewandowska-Tomaszczyk (University of Lodz, Poland)
Email: blt@uni.lodz.pl

Keywords: CMC, dialogue, discourse connection, emotionality patterning, FtF interaction, interactivity patterns, *loose balloons* communication, perception of interpersonal status roles, *ping-pong* communication, *snowball* communication, spatio-temporal conditioning, topic-effect constraints

Introduction

The present study is designed to assess the contribution of dialogic elements to synchronous dyadic Computer-Mediated Communication (CMC) as compared with Face-to-Face (FtF) Interaction and, secondly, to identify the basic contrasts between English and Polish in this respect. The elements analysed are (a) spatio-temporal conditioning of communication, (b) interactivity patterns in FtF and CMC, (c) topic-effect constraints, (d) discourse connection, (e) emotionality patterning, and (f) perception of interpersonal status roles.

The analysis involves samples of spoken conversational components of British National Corpus (BNC) and National Corpus of Polish (nkjp.pl) and the PELCRA corpora of Polish and English internet materials (PELCRA, Institute of English, University of Lodz). The internet language analysed comprises comments on online newspaper articles and to youtube presentations from the years 2011-2014.

The research methods applied are qualitative and quantitative. They include a study of the use of particular *discourse strategies* and *their linguistic realization* analyzed in terms of the type/token ratio, forms of address, metaphor and other figurative uses, utterance positive or negative polarity, axiology status expressed as valence associated with particular judgments and opinions, and their verbal manifestations. These occurrences are studied in terms of their frequencies of occurrence relative to dialogue topic, length and communication type.

Spatio-temporal conditioning

Both in CMC and FtF interaction the exchange is synchronous. The samples used for FtF discourse analysis are dyadic conversations, while in the case of CMC communication the general pattern are many-to-many exchanges, with frequent one-to-one (local) interactions between two individual users which are studied as materials in the present paper.

Typology of CMC and FtF Communication

Three types of online discourse practices involving comments on online newspaper articles are identified in Lewandowska-Tomaszczyk (2014):

1. 'Ping-pong' communication between two, usually individual, interaction participants represents an autonomous and confrontational profile, targeted towards two polar judgments. It includes a relatively high number of feedback loops (replies, *likes*). It is framed in an argumentative, aggressive, discussion type.
2. 'Snowball' communication has a fully determined communicative profile with a clearly defined ultimate objective and an external opponent. The moves and turns are equally or more strongly confrontational than in the ping-pong type. The structure has an observable magnifying axis - stimulated by an increasing flow of energy present and rising, which leads to a climax, and not infrequently success, in real life (as e.g. in ACTA and OCCUPY movements).
3. 'String balloons' communication presents a looser interactional structure often around issues of social and moral value. This communication practice is weakly polarized and contains no one climax. It represents rather sets of interactional moves back and forth along a controversial theme, with frequent reminiscence of the individual's personal life experiences, in which users often digress from the main topic of the exchange.

It is argued in the present paper that some of the types (particularly ping-pong and loose-balloon types) correspond to the interactional dynamics of FtF discourses whereas some others (snowball communication) are shown to be more constrained and predominantly occurring in online many-to-many mass-communication exchanges. Furthermore, some of the relevant sub-types are preferably used in particular culture/language-bound contexts. The ping-pong type in both languages is schematic of a strongly confrontational two-party exchange:

National Corpus of Polish (dialogic)

A: namawiasz ją, by nie grała w drużynie narodowej Lit. 'you've been persuading her not to play in the national team'

B: - Bzdura - Jaki miałbym w tym interes?! Mnie zależy na tym, aby właśnie grała, wtedy wartość zawodniczki idzie ostro w górę. Lit. 'Rubbish! What interest should I have in it? I just care for her to play, as then the player's value sharply increases'

Polish (internet comments)

Offensive, often vulgar ping-pong is more frequent in Polish CMC (Lewandowska-Tomaszczyk in press)

A: A myślałem, że piłka nożna to gra dla ciot. 'And I thought that football is a game for gay prople (offensive)'

B: a co? szukasz sportu dla siebie? 'Why, are you looking for a sport for yourself?'

A: oho, widzę, że kolejny siatkarz/piłkarz się obraził :) 'oh, I see that another volleball/football player feels offended'

English

Ping-pong (internet – dialogic)

- Every single country in the EU needs a referendum; the people had no say, their governments joined in whether the people wanted it or not. Cameron's motive may be political but a good one anyway.

- Sorry, but clearly no. Where does this unflinching believe into the wisdom of referendums come from? I really feel the idea of "referendum" is almost exclusively brought up by those who realize their position on a single question has no representative majority.

Loose balloon (internet - dialogic)

A: I am a West Ham fan. When we are losing 4-0 with seconds left we sing: "5-4! we're gonna win 5-4!" It's an exquisite moment of gallows humour which, as you can imagine, I have enjoyed many times. To cut it short by blowing the whistle early is cruelty beyond words.

B: Ah, West Ham... wasn't there a football club that went by that name once? Such memories.

Topics

Although most of the topics discussed in FtF interaction overlap with those in CMC and present people, events or opinions, with a varying degree of reference to the commentators' individual lives and experiences, FtF interactions (private, non-surreptitious) are rarely observed to lead to effects on a global scale, present in the CMC snowball communication type.

Discourse Markers

The analysis identifies a number of discourse connectors which are used mainly in spoken FtF conversation and convey mainly negative meanings such as English modal-volitional-evaluative *Why x?* Pol. *Dlaczego x?*, *Oh no!* Polish *no nie!* (Lewandowska-Tomaszczyk 2004, Lewandowska-Tomaszczyk & Tomaszczyk 2014), English *Not that*, Pol. *Nie (to), że(by)* (Schmid 2013) in some of their functions. In spoken Polish, and less often in CMC, a range of discourse connectors introduced by the particle *no* is used. The meaning of Polish *no* corresponds to a number of English sentence-connecting senses and approximates 'well / then/ all right' as in: *(No) tak* 'well yes', (*No) właśnie* 'just/precisely', *(No) dobra* 'well, all right'.

Emotionality

Strongly negative emotionality patterns are much more frequent in Polish CMC in the dyadic ping-pong exchange type (Lewandowska-Tomaszczyk in press), and less frequent in English,

although this relation appears topic-sensitive (e.g. British presence in the European Union Lewandowska-Tomaszczyk 2013). Language-specific emotionality patterns are generated in the present work on the basis of the type and frequency of emotionally charged utterances, phrases and words.

Interpersonal status role perceptions

It is observed in both the CMC and FtF data that interactants perceive the hierarchical dominance of interpersonal roles in both languages, although in Polish CMC the interpersonal roles appear to play a less important part. Although this finding partly supports those research proposals which assume the presence of the impoverished social cues in CMC, the tendency is not seen to be universal (Ziegele 2014 for a discussion of dominant/subordinate roles in dyads).

Conclusions

Results comparing face-to-face (FtF) and synchronous CMC dialogues in cross-linguistic contacts indicate both inter-modal as well as cross-linguistic/cultural differences with respect to communicative preferences, as reflected in the investigated language structures. Worth further investigation is the amount and role of confrontational, negative meanings present in CMC dialogues, and asymmetrically distributed in English and Polish.

Acknowledgement: Research carried out within COST Action TD0904 TIMELY, supported by National Science Centre (NCN) grant No 2011/01/M/HS2/03042, *Perception of Time as a Linguistic Category*.

References

boyd, Danah, 2001. *FACETED ID/ENTITY: Managing representation in a digital world.* A.B. Computer Science. Brown University. Providence, Rhode Island.

Lewandowska-Tomaszczyk, Barbara (in press) In: *Languages, Cultures, Media*. Ed. By Barbara Lewandowska-Tomaszczyk, Monika Kopytowska, John Osborne. Josef Schmied, Konca Yumlu.

Lewandowska-Tomaszczyk, Barbara. (2014). "Emergent Group Identity Construal in Online Discussions: A Linguistic Perspective". In: *Revitalizing Audiences: Innovations in European Media Research.* Ed. by Frauke Zellen, Cristina Ponte, Brian O'Neill. Routledge.

Lewandowska-Tomaszczyk, Barbara. 2013. Online Interconnectivity and Negative Emotion Patterning. Special issue of *"Sociedad de la Información"*. In: *New media, audience and emotional connectivity*. Ed. by Hada M. Sánchez Gonzales, 76-109.

Lewandowska-Tomaszczyk, Barbara 2004. Conceptual blending and discourse functions. The case of *no nie 'oh, no'*. *Research in Language* 2. 33–47.

Lewandowska-Tomaszczyk, Barbara & Jerzy Tomaszczyk. 2014. "Negative meanings in Polish and English event reference in discourse: a cognitive corpus-based study" submitted.

Schmid. Hans-Joerg 2013. Is usage more than usage after all? The case of English *not that.* Linguistics 1, 51: 75-116.

Ziegele, Marc (2014)"Differences between online user comments and traditional news conversations". Paper given at the final COST Action meeting *Transforming audiences, transforming societies*, Ljubljana.

Samples

British National Corpus

National Corpus of Polish nkjp.pl

Internet comments to online publications

UK *Telegraph*, European Football Championship EURO2012 finals between Spain and Italy on 3 July 2012 **(English)** http://blogs.telegraph.co.uk/news/tomchiversscience/100168696/euro-2012-iker-casillas-spain-and-sportsmanship

Volleyball player of the winning team in a match between Trentino and Rome *Osmany Juantorena's Funny Serve Insults Opponents* **(Polish)**http://www.youtube.com/watch?v=9YgyXFOZnMs&feature=player_embedded

UK EU Referendum: Huffington Post (Reuters) UK EU "Referendum: David Cameron Promises In-Out Vote In 2015' by Andrew Osborn and Peter Griffiths, published on 23 January 2013 **(English)**

Studying the Effects of Affective Feedback in Embodied Tutors

Mei Yii Lim, Mary Ellen Foster, Srini Janarthanam,
Amol Deshmukh, Helen Hastie and **Ruth Aylett**
School Of Mathematical and Computer Science
Heriot-Watt University,
EH14 4AS, Edinburgh, Scotland, UK
m.lim@hw.ac.uk

Abstract

We present a study designed to explore the effect of feedback on perception of an interactive embodied agent as well as the overall performance and experience of primary school children aged 12-13 carrying out a treasure hunt activity. We use an interactive dialogue agent to compare three experimental conditions: no feedback, neutral feedback, and affective feedback. We study if the agent in the affective condition helps in engagement of the task more than the two other conditions.

1 Introduction

Emotions play an important role in human-human interaction (Damasio, 1994). Agents that exhibit human-like emotions have now become a commonplace in the domain of human-computer interaction. Starting from the pioneering work of (Bates, 1994) and (Picard, 1997), emotional agents now exists in various applications to serve different purposes including but not limited to military (Gratch and Marsella, 2004), health (Bickmore and Picard, 2005), commerce (Gong, 2007), tourism (Lim, 2007), video games (Isbister, 2006) and education (Okonkwo and Vassileva, 2001; Prendinger et al., 2003; Dias and Paiva, 2005; Maldonado et al., 2005). In education, emotional expressions have been incorporated into embodied teaching agents with the aim of improving learning experience in users. Although inclusion of emotional expressions into virtual tutors rarely lead to negative interaction, positive effect was not always achieved on learning experience (Beale and Creed, 2009).

In this paper, we present an experiment to investigate how feedback—none, neutral, or affective—affects a child's perception, experience and performance in a real-world treasure hunt activity. This work takes place in the context of the EU project EMOTE[1] (EMbOdied-perceptive Tutors for Empathy-based learning) which aims to develop virtual tutors that have the perceptive and expressive capabilities to engage in empathic interactions with learners in school environments, grounded in psychological theories of emotion in social interaction and pedagogical models for learning facilitation.

2 The Treasure Hunt

2.1 The Experiment

The treasure hunt activity requires a child to apply his/her map reading skills and is aimed at primary school children aged 12-13. There will be three experimental conditions: no feedback, neutral feedback and affective feedback. In the no feedback condition, students will be given paper maps and instructions, and will not interact with an embodied agent at all during the treasure hunt. In the other two conditions, students will be given Android tablets running an application which displays a digital version of the paper map, along with an embodied agent which will present the instructions and pose the questions. This agent will also provide the students with feedback on the correctness of their answers to the questions posed during the treasure hunt; depending on the experimental condition, the feedback will be either neutral or affective.

In total, 36 students will participate in this study. They will carry out the treasure hunt in pairs, resulting in 6 groups per condition. Prior to the treasure hunt, all students will have a short interactive session with a robot called Susie. The robot will introduce the treasure hunt and conduct a short question and answer session to check the students' readiness for the activity. The robot will be controlled by a wizard in the neighbour-

[1] http://www.emote-project.eu/

Figure 1: The Treasure Hunt Application Start Screen

ing room, and will therefore be capable of taking a few questions from the students if necessary. The main aim of this session is to allow the students to interact and familiarise themselves with the robot, which will then appear as an embodied virtual agent on the tablet for the feedback conditions.

Through this treasure hunt activity, we would like to explore the effect of feedback on the students' perception of an embodied agent as well as their overall experience and performance in carrying out the task at hand. In this study we restrict the emotional display to only three basic expressions (neutral, happy and sad) to ensure that the children understand the affective information being communicated.

The feedback includes both emotional facial expressions and utterances. In the affective condition, a happy expression will be displayed accompanied by utterances such as "brilliant, very good, fantastic" when students answer a question correctly, while a sad expression will be displayed accompanied by utterances such as "Oh no, I'm sorry" when they answer incorrectly; in the latter case, the correct answer will also be provided. In the neutral condition, the agent will always display a neutral expression and reply with "correct" or "incorrect" utterances.

2.2 Treasure Hunt Application

We have designed and implemented a treasure hunt Android application for the above study. In order to compare the three experimental conditions, we have kept the features of the application to be as close to the paper version as possible, except for the addition of the embodied character Susie.

Each step starts with the virtual character presenting a task and questions to the user through speech. Subtitles are displayed on screen in case the students missed what Susie was saying, and the students can also replay the speech at any point if necessary. Each task requires the students to walk a few yards making use of their map skills. At the end of each walk, the students have to confirm their arrival.

The system will then re-present relevant questions related to the task with multiple choice answers and the students are required to select an answer from the given choices. Depending on whether the answer is correct or not, the system responds with appropriate feedback: neutral or affective. In the paper version, the students are also presented with multiple choice answers of which they have to circle the correct one.

2.3 Data Collection

Following the treasure hunt, the students will answer a short questionnaire. It focuses specifically on the children's perception of the embodied agent and their overall experience of the treasure hunt activity, applying the combination of Godspeed likeability items (Bartneck et al., 2009) and the Smileyometer, an instrument used to measure enjoyment and fun (Read and Macfarlane, 2002) aiming to make the task of answering the questionnaire more interesting for the target group. The Smileyometer uses pictorial representations of different kinds of happy faces to depict the diverse level of satisfaction according to 5-point Likert scale.

3 Conclusion and Future Work

By the time of this workshop we will have analysed and deduced reasonable answers to our research questions which hopefully will provide insights to our future design of an empathic tutor.

Acknowledgements

This work was partially supported by the European Commission (EC) and was funded by the EU FP7 ICT-317923 project EMOTE. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

References

C. Bartneck, E. Croft, and D. Kulic. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81.

Joseph Bates. 1994. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, Jul.

R. Beale and C. Creed. 2009. Affective interaction: How emotional agents affect users. *Human-Computer Studies*, 67:755–776.

T. Bickmore and R. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer Human Interaction (TOCHI)*, 12(2):193–327.

Antonio Damasio. 1994. *Descartes' Error: Emotion, Reason and the Human Brain*. Gosset/Putnam Press, New York.

J. Dias and A. Paiva. 2005. Feeling and reasoning: A computational model for emotional agents. In *12th Portuguese Conference on Artificial Intelligence (EPIA 2005)*, pages 127–140, Portugal. Springer.

L. Gong. 2007. Is happy better than sad even if they are both non-adaptive? effects of emotional expressions of talking-head interface agents. *International Journal of Human Computer Studies*, 65(3):183–191.

J. Gratch and S. Marsella. 2004. A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306.

K. Isbister. 2006. *Better Game Characters by Design: A Psychological Approach*. Morgan Kaufmann.

M. Y. Lim. 2007. *Emotions, Behaviour and Belief Regulation in An Intelligent Guide with Attitude*. Ph.D. thesis, School of Mathematical and Computer Sciences, Heriot-Watt University, Ediburgh, Edinburgh.

H. Maldonado, J.R. Lee, S. Brave, C. Nass, H. Nakajima, R. Yamada, K. Iwamura, and Y. Morishima. 2005. We learn better together: enhancing elearning with emotional characters. In T. Koschmann, D. Suthers, and T.W. Chan, editors, *Computer Supported Collaborative Learning 2005: The Next 10 Years!*, pages 408–417. Lawrence Erlbaum Associates, Mahwah, NJ.

C. Okonkwo and J. Vassileva. 2001. Affective pedagogical agents and user persuasion. In C. Stephanidis, editor, *Proceedings of the 4th International Conference on Universal Access in Human Computer Interaction*, pages 5–10, Beijing, China. Springer.

R. W. Picard. 1997. *Affective Computing*. MIT Press.

H. Prendinger, S. Mayer, J. Mori, and M. Ishizuka. 2003. Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces. In T. Rist, R. Aylett, D. Ballin, and J. Rickel, editors, *Fourth International Working Conference On Intelligent Virtual Agents (IVA 03)*, pages 283–291, Kloster Irsee, Germany. Springer.

Janet Read and Stuart Macfarlane. 2002. Endurability, engagement and expectations: Measuring children's fun. In *Interaction Design and Children, Shaker Publishing*, pages 1–23. Shaker Publishing.

Towards Deep Learning for Dialogue State Tracking Using Restricted Bolzman Machines and Pretraining

Callum Main, Zhuoran Wang, and Verena Rieser
Interaction Lab
Heriot-Watt University
Edinburgh, UK
www.macs.hw.ac.uk/InteractionLab

Abstract

Dialogue state tracking aims to estimate the user's goal over the course of a dialogue. Recently, deep neural networks have shown to be successful in this task, especially for generalising to unseen states. In this research, we investigate an alternative deep learning framework, using Restricted Bolzman Machines with pretraining. We aim to show that, by adding a pre-training phase which allows to initialise learning from unlabelled data, leads to significant improvements in terms of accuracy over a baseline using Deep Neural Networks with shared initialisation.

1 Introduction

Statistical spoken dialogue systems maintain a distribution over possible dialogue states ("belief state") in order to correctly estimate the user's true goal, while communicating with a user, from a noisy and often ambiguous input signal. This process is called *dialogue state tracking* (DST).

Deep neural networks (DNN) have shown to be successful in capturing error correlations for improving Automatic Speech Recognition (ASR) systems, e.g. (Deng et al., 2013), and have recently shown successful for dialogue state tracking (Henderson et al., 2013; Henderson et al., 2014). In this research we extend this previous work by (Henderson et al., 2013) in using Restricted Bolzman Machines (RBMs) with pretraining for initialisation (Hinton and Osindero, 2006), which allows us to utilise an additional data set of 10,619 unlabelled calls to capture correlated hypotheses.

2 Related Work

A recent paper by (Henderson et al., 2013) describes a DNN approach to DST using a simple feed-forward architecture with three layers, see Figure 1. The input consists of feature functions, $f_i(t, v)$ which extract information related to the SLU hypotheses, as well as the machine acts at a particularly turn t. The input layer then consists of the feature functions being summed for turns $t-T$ where T is the window size that has been chosen. This input layer is then combined with three hidden layers, each consisting of a weight matrix $\mathbf{W}_i$ and a bias vector $\mathbf{b}_i$, these layers are then reduced to a single node $E(t, v)$. The overall distribution of the tracker is then given by:

$$P(s = v) = e^{E(t,v)}/Z;$$
$$P(s \notin S_{t,s}) = e^{B}/Z;$$
$$Z = e^{B} + \sum_{v' \in S_{t,s}} e^{E(t,v')};$$

The model was then trained using Stochastic Gradient Descent (with mini-batches to speed up computation) using three different initialisation schemes: training a model for each slot, training a model independent of slot, and training a slot independent model for a few epochs before switching to separate models for each slot. An experimental evaluation revealed that shared initialisation, where a single model is trained first before training separate models for each slot, performed the best. This result leads to using more efficient deep learning techniques such as stacked RBMs using pre-training (Hinton and Osindero, 2006) as a promising direction for research.

3 RBM with Pretraining

Training DNNs is very computationally expensive. As such, recent work uses more efficient models such as deep belief networks (DBNs). A DBN can be efficiently trained in an unsupervised, layer-by-layer manner where the layers are typically made of restricted Boltzmann machines (RBMs). RBMs are stochastic generative artificial neural networks which learn a relationship between a set of visible input units and hidden units

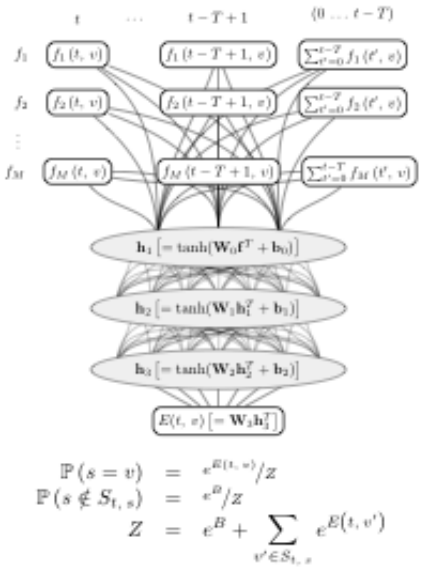


Figure 1: Neural Network structure for computing $E(t, v)$ from (Henderson et al., 2013)

in the form of weighted connections. We can then use an energy based probability model to define a probably distribution that is given by:

$$E(v,h) = -\sum_{i \in visible} a_i v_i - \sum_{j \in hidden} b_i h_j - \sum_{i,j} v_i h_j w_{ij};$$

where v_i, v_j are the states of the hidden unit i and hidden unit j, a_i, b_j are their biases and w_{ij} is the difference between them. The probability for each possible pair of hidden and visable units is then given by:

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\, e^{-E(\mathbf{v},\mathbf{h})};$$

Where Z is the partition function given by summing over all of the possible visable-hidden pairs:

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})};$$

We can then find the probability of assigning to a visible vector by computing the marginal probability by summing over all of the possible hidden vectors:

$$p(\mathbf{v}) = \frac{1}{Z} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})};$$

The derivative of the log likelihood with respect to can then be written as:

$$\frac{\partial P(v)}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model};$$

(Hinton and Osindero, 2006) discovered that RBMs can be trained layer-wise using unsupervised pretraining and then stacked together to great a deep neural network. The pretraining algorithm is then:
(1) Train an RBM using the input X as the visible layer to be used as the first layer.
(2) Transform X using the first layer to obtain data for second layer by either sampling or computing mean activation of hidden units.
(3) Repeat steps 2 and 3 for the desired number of layers.

4 Data

We train and evaluate our approach on data available as part of the first Dialogue State Tracking Challenge (DSTC1) (Williams et al., 2013), see table 4. Results from DSTC1 show that Henderson et al.'s (2013) model outperforms most other models on test set 4, showing its ability to generalise to unseen states. We aim to improve over these results by taking advantage of training set 1b and 1 c, which contain 10,619 unlabelled calls for pretraining.

Set	**no. calls**	**Notes**
train 1a	1013	Labelled training data.
train1b&c	10619	Some SDS as train 1a, but **unlabelled**.
train2	678	Similar to train1.
train3	779	Different SDS to other data sets.
test1	765	Very similar to train1 and train2.
test2	983	Somewhat similar to train1 and train2.
test3	1037	Very similar to train3.
test4	451	unseen Spoken Dialogue Systems.

Table 1: Data released through DSTC1 (Williams et al., 2013)

5 Summary and Future Work

This paper describes work in progress towards a more efficient model for training neural networks for dialogue state tracking using Restricted Bolzman Machines with pretraining, following (Hinton and Osindero, 2006). We compare this model against Deep Neural Networks with shared initialisation as proposed by (Henderson et al., 2013). Full results will be presented at the poster session of SemDial 2014. Having a more efficient way of training while making use of unlabelled data, will allow us to investigate more complex models, for example to directly estimate the dialogue state based on ASR output rather than SLU hypothesis (Henderson et al., 2014).

References

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero. 2013. Recent advances in deep learning for speech research at microsoft. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

M. Henderson, B. Thomson, and S. J. Young. 2013. Deep Neural Network Approach for the Dialog State Tracking Challenge. In *Proc. of SIGdial Workshop on Discourse and Dialogue*.

M. Henderson, B. Thomson, and S. J. Young. 2014. Word-based State Tracking with Recurrent Neural Networks. In *Proceedings of SIGdial*.

Geoffrey E. Hinton and Simon Osindero. 2006. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006.

Angeliki Metallinou, Dan Bohus, and Jason D. Williams. 2013. Discriminative state tracking for spoken dialog systems. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Referential Grounding for Situated Human-Robot Communication

Vivien Mast, Daniel Couto Vale, Zoe Falomir, Mohammad Fazleh Elahi
SFB/TR-8 Spatial Cognition
University of Bremen
{vivien.mast, danielvale}@uni-bremen.de,
{zfalomir, fazleh}@informatik.uni-bremen.de

Abstract

We present a dialogue system and reference handling component for efficient and natural referential grounding dialogues from 2D images. Using a probabilistic representation of qualitative concepts, the system uses flexible concept assignment in reference handling for bridging conceptual gaps between the system and the user, and engages in clarification dialogues based on an evaluation of miscommunication risk.

1 Introduction

From her comfortable sofa, Mary asks her personal assistant robot Amanda: *Could you pass me that yellow book on my desk?* Amanda is not sure which book Mary meant, and asks: *Do you mean the one in front of the coffee cup?* Slightly annoyed, Mary replies: *No, not the green one, the yellow one.* Amanda confirms: *Oh, ok. I thought that was orange. I'll get it.* and brings the book to Mary.

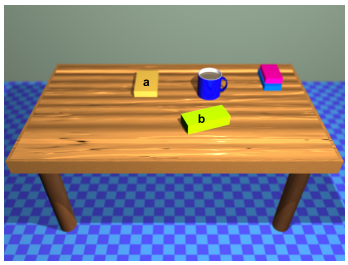


Figure 1: Intended referent (a) and distractor (b).

While standard algorithms for referring expression generation (REG) assume that objects can be defined by a fixed set of crisp properties (Krahmer and van Deemter, 2012; Dale and Reiter, 1995), humans carve up the world in multiple ways (Steels, 2008), depending on sensory differences (Roorda and Williams, 1999), exposure to a domain (Goldstone et al., 2012, p. 621), or situational conditions (Spranger and Pauw, 2012).

In order to bridge conceptual gaps between interactants and establish common ground, humans flexibly adapt the use of concepts (Clark and Brennan, 1991; Garrod and Anderson, 1987; Clark and Wilkes-Gibbs, 1986). As the conceptual gap in human-machine interaction is even larger, dialogue systems may benefit greatly from sophisticated grounding abilities.

Our implementation of the agent based dialogue system architecture and framework DAISIE (Ross and Bateman, 2009), using the ***P**robabilistic **R**eference **A**nd **GR**ounding* mechanism PRAGR (Mast and Wolter, 2013b; Mast and Wolter, 2013a), flexibly assigns properties during reference handling in order to maximize communicative success. We show how the DAISIE+PRAGR system is capable of engaging in grounding dialogues about images as they may be provided by a camera installed on the head of a mobile robot, by generating and resolving referring expressions (REs), and using probabilistic evaluations of the REG and reference resolution (RR) output for making reasonable dialogue decisions.

2 PRAGR

PRAGR is a probabilistic reference handling system for enabling dialogic grounding, described in detail by Mast and Wolter (2013b). PRAGR's core concepts are *acceptability*—the probability $P(D|x)$ that the interlocutor accepts D as a description of object x, and *discriminatory power*—the probability $P(x|D)$ that D discriminates x from its distractors, a value comparing the acceptability of D for the target to its acceptability for distractors. The stochastic model handles descriptions of arbitrary complexity, including relations.

In RR, given a description D, PRAGR selects as best referent x^* the object for which D has the highest acceptability. In REG, PRAGR aims for

effectiveness in communication, given uncertain knowledge. Thus, it searches for the most *appropriate* description D^* which jointly maximizes acceptability and discriminatory power.

3 Layered Representation

In PRAGR's two-layer knowledge representation, the *perceptual layer* represents qualitative and metric perceptual properties of objects (e.g. defining shape points and hue, lightness and saturation) obtained in a first step of abstraction (Falomir et al., 2012). *Perceptual grounding* modules provide probabilistic mappings of objects to conceptual properties such as ORANGE in the *conceptual layer*. A *dialogic grounding module* may add or overwrite mappings on the conceptual level.

Perceptual grounding modules include a probabilistic model of projective terms adapted from Mast and Wolter (2013b), a crisp model of object type based on Qualitative Image Description (Falomir et al., 2012) and a fuzzy adaptation of the colour model by Falomir et al. (2013).

With these probabilistic feature models, PRAGR can consider gradual differences in discriminatory power and acceptability and provide the most appropriate description. It may call the same ball *the red ball* (Figure 2a) or *the orange ball* (Figure 2b), depending on present distractors, as acceptability of a property for distractors dampens discriminatory power.

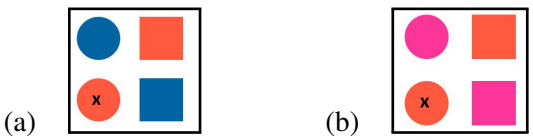


Figure 2: Context sensitivity of category assignment: (a) the red circle, (b) the orange circle.

4 DAISIE+PRAGR

The current update cycle of DAISIE's information state depends on 5 subprocesses for the automation of linguistic understanding and on 4 subprocesses for the automation of linguistic expression as shown in Figure 3. REs in the input are identified during experiential interpretation and enriched via co-reference resolution against the discourse history during textual interpretation. They are then queued for handling by the dialogue manager which directly accesses PRAGR. In the following dialogue: **H**uman: *Bring me the red box.*,

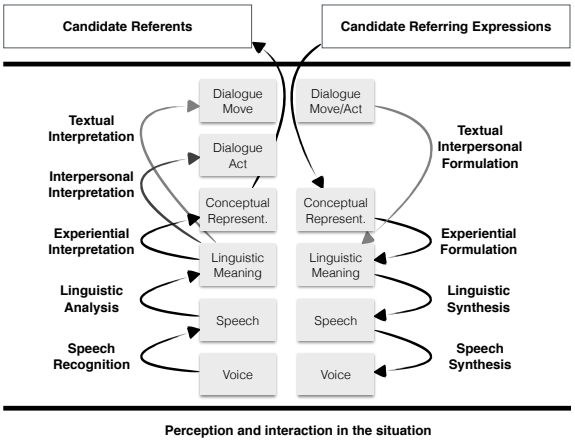


Figure 3: Architecture of DAISIE

System: *Which red box do you mean?*, H: *The one on the floor.*, the expression *one* is resolved to the conceptual representation [RED, BOX, SUPPORT(FLOOR)] before being passed to PRAGR which then provides an n-best list of potential referents with the evaluation values of the input expression. Based on this evaluation, the dialogue manager plans the next dialogue move. For example, if there is no substantial difference between candidates, an open clarification question (*Which red box do you mean?*) is generated. If one preferred candidate is found, depending on acceptability and discriminatory power, the system may generate an expansion (*Do you mean the one on the table?*) or a confirmation (*OK, I'm getting it.*).

In generation, conceptual representations of REs are selected and evaluated by the reference handling component, called as part of the dialogue move selection. If no sufficiently appropriate RE for an intended target can be found, the system may attempt to ground a potential reference object first, in order to use this for a follow-up reference to the intended target: S: *Can you see the low table to the left of the door?* H: *Yes.* S: *Your keys are in the small green box on that table.*

5 Summary

The proposed referential grounding dialogue system DAISIE+PRAGR is capable of flexibly using concepts in order to improve referential success in generation and understanding. Decisions of the dialogue manager about next dialogue moves are informed by the evaluation results of the reference handling component, thus enabling natural and efficient grounding dialogues in situated communication.

Acknowledgments

Funding by the Deutsche Forschungsgemeinschaft (DFG) for the SFB/TR 8 Spatial Cognition, project I5-[DiaSpace] and the European Commission through FP7 Marie Curie IEF actions under project COGNITIVE-AMI (GA 328763) is gratefully acknowledged.

References

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition*, 13:127–149.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1 – 39.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

Zoe Falomir, Lledó Museros, Gonzalez-Abril Luis, M. Teresa Escrig, and Juan A. Ortega. 2012. A model for the qualitative description of images based on visual and spatial features. *Computer Vision and Image Understanding*, 116(6):698–714.

Zoe Falomir, Lledo Museros, Luis Gonzalez-Abril, and Ismael Sanz. 2013. A model for qualitative colour comparison using interval distances. *Displays*, 34:250–257.

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

Robert L. Goldstone, Alan Kersten, and Paulo F. Carvalho. 2012. Concepts and categorization. In A. F. Healy and R. W. Proctor, editors, *Comprehensive handbook of psychology, Volume 4: Experimental psychology*, pages 607–630. Wiley, New Jersey.

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

Vivien Mast and Diedrich Wolter. 2013a. Context and vagueness in REG. In *Proceedings of PRE-CogSci 2013*, December.

Vivien Mast and Diedrich Wolter. 2013b. A probabilistic framework for object descriptions in indoor route instructions. In T. Tenbrink, J. Stell, A. Galton, and Z. Wood, editors, *Spatial Information Theory*, volume 8116 of *Lecture Notes in Computer Science*, pages 185–204. Springer International Publishing, October.

Austin Roorda and David R Williams. 1999. The arrangement of the three cone classes in the living human eye. *Nature*, 397(6719):520–522.

Robert J. Ross and John A. Bateman. 2009. Daisie: Information state dialogues for situated systems. In V. Matoušek and P. Mautner, editors, *Text, Speech and Dialogue*, LNCS, pages 379–386, Berlin, Heidelberg. Springer.

M. Spranger and S. Pauw. 2012. Dealing with Perceptual Deviation - Vague Semantics for Spatial Language and Quantification. In L. Steels and M. Hild, editors, *Language Grounding in Robots*, pages 173–192. Springer.

Luc Steels. 2008. The symbol grounding problem has been solved, so what's next? In M. De Vega, A.M. Glenberg, and A.C. Graesser, editors, *Symbols and embodiment: Debates on meaning and cognition*, pages 223–244. Oxford University Press.

Laughter in naturalistic mother-child interaction: from 12 to 36 months

Chiara Mazzocconi [1], Sam Green[1], Ye Tian[2], Caspar Addyman[3]
[1] Division of Psychology and Language Sciences, University College London
[2] Université Paris Diderot
[3] Centre for Brain and Cognitive Development, Birkbeck University of London
chiara.mazzocconi.13@live.ucl.ac.uk

1. Introduction

Laughter is a social vocalization universal across cultures and languages (Ruch & Ekman 2001, Sauter et al. 2010). Rather than being an automatic response to funny stimuli, research shows that it is finely sequentially-organized, timed in interaction and conveys a broad range of meanings even in serious contexts. Laughter emerges as a primitive and unconscious vocalization reflecting positive inner states (Provine, 1996) and, through the modeling and influence of the environment, it becomes an important and deeply social form of non-verbal communication; that is crucial in bonding, establishing relationships and managing interactions. Speakers tend to laugh 46% more than their audience (Provine, 1993) and people are 30 times more likely to laugh when they are not alone, even in the absence of a humorous stimulus (Provine, 2004).

Laughter emerges in infants at around 3 months of age as a response to physical stimulation, and over the first year it is progressively elicited by more and more distal events, e.g. socially inappropriate or incongruous acts (Mireault et al. 2012). This development stems from a marked innate interest in others' actions, emotions and states which support the development of a mind reading ability needed to infer the playful intention of others and to not be scared by incongruous stimuli (Semrud-Clikeman et al. 2010).

Research reports that mothers laugh more frequently and display a higher tendency to join the partner's laughter (i.e. antiphonal laughter) (Table 1) than their children (Nwokha et al., 1994). The rate of infants' laughter, both non-dyadic and antiphonal (Table 1), increases over time becoming correlated with the rate of mothers' laughter (Ziajka, 1981) only by the second year.

The increase of antiphonal laughter can reflect a deeper interest in others' mental states and feelings, a better comprehension of their causes and a higher pleasure in sharing them: meaning e.g. "If you think that this is funny, so do I" (Fogel et al. 1992), which entails a meta-representation of the partner mental state situated in the context in which the laughter occurs.

Non-dyadic Laughter	a laugh not preceded by any laugh from the conversational partner within 4s
Antiphonal (aka dyadic) laughter	a laugh that occurs less than 4s after a laugh by the partner with or without overlap

Table 1: Definitions non-dyadic and antiphonal laughter

Interestingly in children affected by Autism, where social competences, mind reading, empathy and pragmatic skills represent the core of the difficulties, atypical laughter patterns have been reported: despite a typical frequency of laughter, a lower rate of shared laughter, together with the tendency to laugh at inexplicable stimuli, are observed (Reddy et al. 2002).

Laughter emerges in infants long before walking, gesturing and speaking. Laughter behaviour, in terms of frequency, context of occurrence and timing development, may thus serve as an early marker for certain delays or impairments in social, emotional and learning (Bruno et al. 1987) development. It may also be informative and predictive of communicative and language development.

2. Current study

The nature of the current study is in the first instance exploratory, aiming to observe laughter behaviours development in childhood from 12 to 36 months and to investigate the relation to language, communicative and pragmatic abilities, in order to deepen the little research available on the topic. Longitudinal analysis are being carried out on the laughter behaviour of three typically developing native British-English female children. We coded videotapes of natural interactions with their mothers in a familiar context. One of the children showed a slight delay in language acquisition, being possibly labelled as a "late bloomer". Observations will be conducted on two hours of video-recording at five time points for each child: 12, 18, 24, 30, 36 months [1].

We identified the occurrence of laughter both for the child and for the mothers. For each laugh event we specified its context of occurrence, the partner's response, its position in relation to

[1] The videos analysed are part of a larger longitudinal project conducted by Dr. Andrew Nevins and Sam Green investigating the impact of the ambient language on learning speech sounds.

speech (overlapping others' utterance or co-occurence with speech from the speaker) and its temporal sequence parameters: non-dyadic and antiphonal laughter.

Our preliminary results (2 hours of coded video at 5 time points for two children) show a developmental trend in children's explicit response (look, smile or laugh) to mothers' laughter - in contrast to a constant explicit response from the mothers (Figure 1). Most of the laughter occurs in interaction, being elicited more often by action from the partner (physical or verbal) rather than external targets.

The laughter behaviours of the mothers are very different from that of children. Consistent with the literature, mothers laugh more frequently than their children (mean number of laughter occurrences/10 minutes: Mother1: 6.54; Child1: 2.96; Mother2: 6.92; Child2: 1.60), with a rate close to the one reported in adult-adult interaction: 5.8/10min (Vettin and Todt, 2004).

Speech-laughter (i.e. laughter produced simultaneously with speech) is frequently observed in mothers (M1 25%; M2 18%), at a rate very similar to one reported in Nwokah et al. 1999 (18.6%) and contrasting with previous data by Provine (1993) in adult-adult interaction (0.1%). Contrary, the same behaviour is almost absent in children, suggesting that by 36 months, children have not yet developed the ability to integrate laughter into speech. On the other hand, children's laughter never overlaps with partner's speech from at least 12 months, possibly revealing an early acquisition of conversational turn-taking ability.

The percentage of antiphonal laughter is markedly higher for mothers (M1 32%; M2 43%) than children (C1 5%; C2 3%), being close to the percentage observed in adult-adult game-interaction (34%, Smoski and Bachoroski, 2003). Interestingly the rate of antiphonal laughter from the mothers seems to decrease and then become stable as the language ability of the child (as measured by the OCDI[2] and the LINCOLN T-CDI[3]) develops. This is particularly marked for C2 who presented a slight delay in language production: the antiphonal rate by the mother decrease steeply from 12 to 30 months, nearing the percentage of M1 around 30 months, when the language scores of the child reach 50%. These data could indicate a privileged use of laughter as a response to children's laughter when they are not well confident with language.

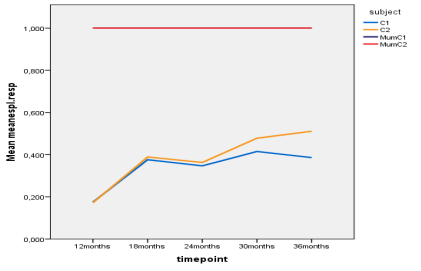


Figure 1: Children (C1, C2) and mothers (M1, M2) explicit responses to partner's laughter.

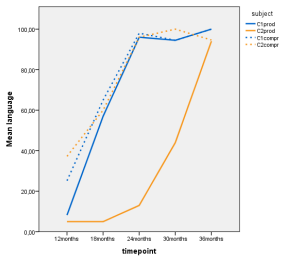


Figure 2: Language inventory scores: production and comprehension (C1; C2).

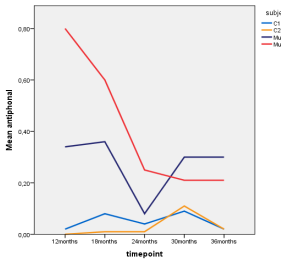


Figure 3: Rate antiphonal laughter overtime by children (C1; C2) and mothers (M1; M2)

3. Conclusion and future research

Overall, our initial results show that children's reactions to mothers' laughter increase from 12 to 36 months. Children as young as 12 months appear to have mastered the conversational turn-taking convention. On the other hand, their antiphonal laughter and speech laughter still sporadic at 36 months. Mothers' laughter behaviours could be influenced by child language ability. Their laughter and antiphonal laughter rate when interacting with an infant are similar to the rates reported in adult-adult interaction, while speech-laughter seem to occur more frequently than in adults' interaction. Ongoing analyses are being conducted on the third child. Data collected at the last time point for empathy, theory of mind and perspective taking ability will be analysed and compared to the observed laughter behaviours. Future research will extend the longitudinal span from 36 months to middle-childhood exploring when adults and children patterns get closer. We will also investigate laughter patterns of mother-child interaction in clinical population in order to better understand the relationship between laughter, language, communicative and cognitive development, contrasting patterns in children affected by Specific Language Impairment and High and Low Functioning Autism. It would be than interesting to investigate whether laughter by mothers decreases over time only in response to laughter or to all non-verbal behaviour by the child.

[2]*Oxford Communicative Development Inventory - a UK adaptation of the MacArthur-Bates CDI*

[3]*Lincoln Toddler Communicative Development Inventory - a UK adaptation of the MacArthur Toddler CDI*

References

Bruno, R. M., Johnson, J. M., & Simon, J. (1987). Perception of humor by regular class students and students with learning disabilities or mild mental retardation. Journal of learning disabilities, 20(9), 568-570

Fogel. A., Nwokah. E., Dedo, J., Messinger, D., Dickson, K.L., Matusov, E., & Holt, S. (1992). Social process theory of emotion: A dynamic systems approach. *Social Development, I,* 122-142.

Mireault, G., Poutre, M., Sargent-Hier, M., Dias, C., Perdue, B., & Myrick, A. (2012). Humour Perception and Creation between Parents and 3- to 6-month-old Infants. Infant and Child Development, 21(4), 338-347

Nwokah, E. E., Hsu, H. C., Davies, P., & Fogel, A. (1999). The Integration of Laughter and Speech in Vocal CommunicationA Dynamic Systems Perspective. *Journal of Speech, Language, and Hearing Research*, *42*(4), 880-894.

Nwokah, E. E., Hsu, H. C., Dobrowolska, O., & Fogel, A. (1994). The development of laughter in mother-infant communication: Timing parameters and temporal sequences. *Infant Behavior and Development*, *17*(1), 23-35

Provine, R. R. (2004). Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science*, *13*(6), 215-218.

Provine, R. R. (1996). Laughter. *American scientist*, 38-45.

Provine, R. R. (1993). Laughter punctuates speech: Linguistic, social and gender contexts of laughter. Ethology 85: 291-298

Reddy, V., Williams, E., & Vaughan, A. (2002). Sharing humour and laughter in autism and Down's syndrome. British Journal of Psychology, 93(2), 219-242.

Ruch, W., & Ekman, P. (2001). The expressive pattern of laughter. *Emotion, qualia, and consciousness*, 426-443.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations.*Proceedings of the National Academy of Sciences*, *107*(6), 2408-2412

Semrud-Clikeman, M., & Glass, K. (2010). The relation of humor and child development: Social, adaptive, and emotional aspects. *Journal of child neurology*, *25*(10), 1248-1260.

Smoski, M., & Bachorowski, J. A. (2003). Antiphonal laughter between friends and strangers. *Cognition & Emotion*, *17*(2), 327-340.

Vettin, J., & Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, *28*(2), 93-115.

Ziajka, A. (1981). *Prelinguistic communication in infancy. New* York: Praeger.

Initiative Patterns in Dialogue Genres

Angela Nazarian and Elnaz Nouri and David Traum
USC Institute for Creative Technologies
12015 Waterfront Dr
Playa Vista, CA 90094, USA
{nazarian,nouri,traum}@ict.usc.edu

1 Overview

One of the ways of distinguishing different dialogue genres is the differences in patterns of interactions between the participants. Morbini et al (2013) informally define dialogue genres on the basis of features like user vs system initiative, amongst other criteria. In this paper, we apply the multi-label initiative annotation scheme and related features from (Nouri and Traum, 2014) to a set of dialogue corpora from different domains. In our initial study, we examine two question-answering domains, a "slot-filling" service application domain, and several human-human negotiation domains.

2 Dialogue Domains

The Twins are two life-size virtual characters who serve as guides at the Museum of Science in Boston (Swartout et al., 2010). The characters promote interest in Science, Technology, Engineering and Mathematics (STEM) in children between the ages of 7 and 14. They are question-answering characters, but unlike SGTs Blackwell and Star, the response is a whole dialogue sequence, potentially involving interchange from both characters, rather than a single character turn.

Amani (Artstein et al., 2009) is an advanced question-answering character used as a prototype for systems meant to train soldiers to perform tactical questioning.

Radiobots (Roque et al., 2006) is a training prototype that responds to military calls for artillery fire in a virtual reality urban combat environment. This is a domain in the slot-filling genre, where there is a preferred protocol for the order in which information is provided and confirmed. Users are generally trainees, learning how to do calls for fire, they are motivated users with some training.

Farmer's Market Negotiation (Carnevale, 2013) are bilateral role-play negotiations between undergraduate business students. The owners of the two restaurants had asked the participants to go to the market and get some apples, bananas, lemons, peppers and strawberries. Each participant has a different payoff matrix for the value of items, and the goal of the negotiation is to partition the items. Initiative annotations for this dataset were used in (Nouri and Traum, 2014)

Cartoon Negotiation (Ziebart et al., 2012) are role-play negotiation dialogues in which two participants negotiate on several issues, each of which has several possible values, and the payoff matrix for the issues differs between the participants.

3 Initiative Annotation Scheme

We use the initiative annotation scheme from (Nouri and Traum, 2014). This scheme breaks both *initiative* and *response* into two distinct concepts, for 4 label, total. For initiative, first there is providing unsolicited, or optional, or extra material, that is not a required response to a previous initiative (N for new). Second, there is the sense of putting a new discourse obligation (Traum and Allen, 1994) on a dialogue partner to respond (I for Invoke obligation). These two concepts often come together, such as for new questions or proposals that require some sort of response: they are both unsolicited and impose an obligation, however, it is also possible to have each one without the other. Statements can include new unsolicited material, without imposing an obligation to respond (other than the weak obligation to ground understanding of any contribution). Likewise, clarification questions impose new obligations on the other, but often do not contribute new material or are not optional, in that the responder can not reply appropriately without the clarification. For response, one

concept concerns fulfilling obligations imposed by prior initiatives (labelled F, for fulfillin obligation). To not do so could be considered rude and a violation of conversational norms in some cases. This is only relevant, if there is an existing initiative-related obligation as part of the conversational state. Another concept generalizes the notion of response to anything that contributes to the same topic and makes an effort to relate to prior utterances by the other party, whether or not it fulfills an obligation or whether there even is a pending obligation (labelled R for related).

4 Initial Results

We have so far annotated at least five dialogues in each of the domains in 2. We are analyzing several automatically extracted features based on the label, including

- the count of each label (I,F,R,N) per negotiation and per person
- the ratio, difference and absolute difference of the number of labels for each person against the number of labels for their negotiation counterpart
- the above measures normalized by the number of turns in dialog
- **Within-turn patterns** the number of all possible combinations of labels for each utterance. There are 16 possible combinations for the 4 types of labels that can be shown as tuples (R,F,I,N).
- **Across-turn Patterns** the number of all possible sequences of labels across two adjacent turns. There are also 16 possible combinations capturing how often each label is followed by labels.

Preliminary findings show differences on a number of dimensions. As expected, the Twins domain in the simple question-answering genre had the highest percentage of F annotations overall, and a disparity between user (many I's) and twins (many F's). The Amani domain was broadly similar, though included a higher percentage of total I's, given that the character often responded to questions with offers or grounding moves. Negotiation domains tended to be more symmetric amongst the distribution of moves to participants.

We intend to continue this annotation and analysis and present more complete results at the workshop.

Acknowledgments

References

Ron Artstein, Sudeep Gandhe, Michael Rushforth, and David R. Traum. 2009. Viability of a simple dialogue act scheme for a tactical questioning dialogue system. In *DiaHolmia 2009: Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue*, page 43–50, Stockholm, Sweden, June.

Peter J Carnevale. 2013. Audio/video recordings of bilateral negotiations over synthetic objects on a table that vary in monetary value.

Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, Ron Artstein, Dogan Can, Panayiotis Georgiou, Shri Narayanan, Anton Leuski, and David Traum. 2013. Which asr should i choose for my dialogue system? In *Proceedings of the SIGDIAL 2013 Conference*, pages 394–403, Metz, France, August. Association for Computational Linguistics.

Elnaz Nouri and David Traum. 2014. Initiative taking in negotiation. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 186–193, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

A. Roque, A. Leuski, V. Rangarajan, S. Robinson, A. Vaswani, S. Narayanan, and D. Traum. 2006. Radiobot-CFF: A spoken dialogue system for military training. In *Proc. of Interspeech*, Pittsburgh, Pennsylvania, USA.

W. Swartout, D. Traum, R. Artstein, D. Noren, P. Debevec, K. Bronnenkant, J. Williams, A. Leuski, S. Narayanan, D. Piepol, C. Lane, J. Morie, P. Aggarwal, M. Liewer, J. Chiang, J. Gerten, S. Chu, and K. White. 2010. Ada and Grace: Toward realistic and engaging virtual museum guides. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, editors, *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20–22, 2010 Proceedings*, volume 6356 of *Lecture Notes in Artificial Intelligence*, pages 286–300. Springer, Heidelberg.

David R. Traum and James F. Allen. 1994. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

Brian D. Ziebart, Dudik M, Geoffrey Gordon, Katia Sycara, Jayne Adair, and Brent J. 2012. Identifying culture and leveraging cultural differences for negotiation agents. In *Hawaii International Conference on Systems Sciences (HICSS-45)*, January.

Towards Generating Route Instructions Under Uncertainty: A Corpus Study

Verena Rieser and Amanda Cercas Curry
Interaction Lab
Heriot-Watt University
Edinburgh, UK
v.t.rieser@hw.ac.uk

Abstract

The overall aim of this work is to develop a principled data-driven approach for generating route instructions in spatial domains when faced with various types of uncertainty, e.g. user location, view shed, distance to target etc. As a first step, we conduct a corpus study investigating how humans give instructions in different scenarios. We find that human instruction givers produce different route instructions based on the information available to them. This motivates a context-adaptive approach for generating route instructions.

1 Introduction

Systems that generate route instructions have recently attracted a lot of attention from the dialogue and Natural Language Generation (NLG) communities, e.g. (Koller et al., 2007; Dethlefs and Cuayáhuitl, 2011; Dethlefs et al., 2011; Janarthanam et al., 2012; Dräger and Koller, 2012) etc. In this research we investigate how to generate route instructions when faced with uncertainty, e.g. about the user's location, view shed, distance to target etc. As a first step, we conduct a corpus study to empirically investigate how humans give instructions in different scenarios. In particular, we compare object references and quantitative descriptions. Previous research seem to suggest that landmark-based route instructions ("*Walk towards the Castle*") are easier to understand than distance-based ones ("*Walk 300 meters*") (Lovelace et al., 1999; Dräger and Koller, 2012). Here, we investigate the choices human Instruction Givers make when confronted with different types of uncertainty. We draw conclusions based on the different distributions of observed surface forms across three different corpora.

2 Corpus Annotation

We manually annotate navigation instructions in two Wizard-of-Oz corpora collected as part of the SpaceBook project (Janarthanam et al., 2014). We follow an annotation scheme by (Levit and Roy, 2007) developed for the HCRC MapTask corpus (Thompson et al., 1993), which we modify to account for situated dialogues. We also utilise the original annotations from MapTask. These three corpora are collected in different setups and thus introduce different types of uncertainty between Instruction Giver (IG) and Instruction Follower (IF):

MapTask (MT): IF and IG, share the same spatial representation in form of a paper map, i.e. distances and landmarks are known to both. The location of the IF is hidden to the IG.

SpaceBook1 (SB1): The IG follows the IF through the city of Edinburgh while communicating on the phone. That is, the IG knows location and view shed of the IF.

SpaceBook2 (SB2): The IG tracks the IF on Google Maps and also has access to Google StreetView. The exact location of the IF is unknown due to a noisy GPS signal.

The annotation scheme decomposes an utterance into navigational information units (NIUs). These NIUs are then further specified according to various aspects of instruction giving, e.g. actions, path descriptions etc. Here we only report on aspects relevant to generation under uncertainty:

Verification Actions aim to clarify uncertainty about position or orientation of the IF.

Reference Objects serve as anchors for identifying directions or positions.

Quantitative Aspect encode how far the traveler should move.

Label	SpaceBook Example	SpaceBook1	SpaceBook2	MapTask
Total NUIs		316	414	2132
Verif.:Position	<POSITION verifier="WIZARD" reference="LANDMARK">Are you standing outside Informatics in Edinburgh?</POSITION>	17.7%	8.2%	11.02%
Verif.:Orientation	<ORIENTATION verifier="WIZARD" reference="LANDMARK">can you see the National Museum of Scotland in front of you?</ORIENTATION>	4.1%	3.1%	0.8%
Object:Landmark	see above.	43.7%	21.7%	33.1%
Object:Streetname	<POSITION verifier="WIZARD" reference="STREETNAME"> This is West Nicolson Street. </POSITION>	7.6%	22.9%	N/A
Object:Proximity	<TURN descriptor="NIL" reference="NIL" quantitative="PROXIMITY">Turn at the next crossing</TURN>	8.2%	7.0%	N/A
Object:UserCentric	<MOVE descriptor="STRAIGHT" reference="USER" quantitative="FALSE"> Just keep walking in the direction you are going.</MOVE>	17.4%	16.4%	5.6%
Object:Cardinal	<MOVE descriptor="STRAIGHT" reference="CARDINAL" quantitative="FALSE">Please continue walking South. </MOVE>	0.3%	0.2%	31.8%
Object:NIL	<TURN descriptor="LEFT" reference="FALSE" quantitative="FALSE">you wish to turn left</TURN>	17.4%	19.3%	13.3%
Object:Other		5.4	12.5	5.1%
Quantitative:Time	<MOVE descriptor="STRAIGHT" reference="OTHER" quantitative="TIME"> About one minute down the road.</MOVE>	0%	1.2%	N/A
Quantitative:Distance	<MOVE descriptor="STRAIGHT" reference="NIL" quantitative="DISTANCE">follow for three hundred meters. </MOVE>	0%	0.2%	64.8%

Table 1: NUIs label distributions: Frequencies within corpora.

3 Results

We highlight and discuss the main differences observed between the three corpora, based on their frequency of occurrence as summarised in Table 1.

- **Verification actions** make up between 11-22% of all possible actions. They are almost twice as frequent in SpaceBook2 than in SpaceBook1 and MapTask. We hypothesise that this occurred when the IG lost sight of the IF. In general, the IG tends to verify the IF's position, but less so the orientation/ view shed.
- **Landmarks** are the most common reference object in SpaceBook1 and MapTask, but SpaceBook2 uses **Steetnames** more often. This can be attributed to the fact that in this scenario the IG tracks the IF on a digital map where street names are indicated. Whereas in SpaceBook1 Landmarks are more prominent due to the shared view shed.
- **Proximity** is only used in SpaceBook1 and SpaceBook2 since the IG had an estimate of the IF's position, whereas in MapTask the IF's location is hidden. Similarly, **UserCentric** instructions are generated relative to the IF's position and thus only occur in SpaceBook1 and SpaceBook2.
- **Cardinal** directions hardly occur in the SpaceBook scenarios, but in MapTask this information is relative to the paper map, and thus a shared point of reference which is used in over one third of the cases.
- The main difference between SpaceBook and MapTask is the occurrence of **quantitative** descriptions. In the SpaceBook scenarios quantitative descriptions hardly ever occur, whereas in MapTask about 65% of instructions are quantified. We attribute this difference to the fact that distances can be easily estimated from a paper map, whereas distances from a digital map or while walking down the street are harder to judge.

4 Discussion and Future Work

In summary, we find that human instruction givers produce different route instructions based on the information available to them: Landmarks are preferred if the view shed of the instruction follower is known; User centric and instructions based on proximity only occur if the location is known; Cardinal directions occur if the orientation is known; And quantitative descriptions are limited to cases where the scale is known. We therefore conclude that in contrast to claims by previous work, it is not always preferred to generate instructions based on landmarks, but good route instructions depend on the contextual information available. In future work, we will generate context-dependent instructions based on a framework for generation under uncertainty (Rieser and Lemon, 2009; Rieser et al., 2014) and test their effectiveness with real users (Janarthanam et al., 2012). We will also measure inter-annotator agreement and run significant tests for the above annotations.

Acknowledgments

The research leading to this work has received funding from the Engineering and Physical Science Research council, UK (EPSRC) under project no. EP/L026775/1. We would also like to thank Srinivasan Janarthanam and Robin Hill for help and discussion.

References

Nina Dethlefs and Heriberto Cuayáhuitl. 2011. Combining hierarchical reinforcement learning and bayesian networks for natural language generation in situated dialogue. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*.

Nina Dethlefs, Heriberto Cuayáhuitl, and Jette Viethen. 2011. Optimising natural language generation decision making for situated dialogue. In *Proc. of the Annual SIGDIAL Conference on Discourse and Dialogue*.

Markus Dräger and Alexander Koller. 2012. Generation of landmark-based navigation instructions from open-source data. In *Proceedings of the Thirteenth Conference of the European Chapter of the ACL (EACL)*, Avignon.

Srinivasan Janarthanam, Xingkun Liu, and Oliver Lemon. 2012. A web-based evaluation framework for spatial instruction-giving systems. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Srinivasan Janarthanam, Robin Hill, Anna Dickinson, and Morgan Fredriksson. 2014. Proceedings of the eacl 2014 workshop on dialogue in motion. pages 19–27. Association for Computational Linguistics.

Alexander Koller, Johanna Moore, Barbara di Eugenio, James Lester, Laura Stoia, Donna Byron, Jon Oberlander, and Kristina Striegnitz. 2007. Shared task proposal: Instruction giving in virtual worlds. In Michael White and Robert Dale, editors, *Working group reports of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.

Michael Levit and Deb Roy. 2007. Interpretation of spatial language in a map navigation task. *IEEE Trans. Syst., Man, Cybern. B, Cybern*, 37(3).

Kristin L. Lovelace, Mary Hegarty, and Daniel R. Montello. 1999. Elements of good route directions in familiar and unfamiliar environments. In *Proceedings of the International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*, COSIT '99, pages 65–82, London, UK, UK. Springer-Verlag.

Verena Rieser and Oliver Lemon. 2009. Natural Language Generation as Planning Under Uncertainty for Spoken Dialogue Systems. In *Proc. of the Conference of European Chapter of the Association for Computational Linguistics (EACL)*.

V. Rieser, O. Lemon, and S. Keizer. 2014. Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(5):979–994, May.

Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hcrc map task corpus: Natural dialogue for speech recognition. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

SpeechCity: A Conversational City Guide based on Open Data

Verena Rieser, Srinivasan Janarthanam, Andy Taylor, Yanchao Yu and Oliver Lemon
Interaction Lab
Heriot-Watt University
Edinburgh, UK
Email: v.t.rieser@hw.ac.uk; Website: www.macs.hw.ac.uk/InteractionLab

Abstract

We demonstrate a mobile application that assists users in planning a day out within a urban environment. Locality-specific information is acquired from open data sources, and can be accessed via intelligent interaction. For demonstrating the full functionality of the system, we simulate a (user-specified) walking route within the city of Edinburgh, where the system "pushes" relevant information to the user. Through the use of open data, the agent is easily portable and extendable to new locations and domains, which we plan to demonstrate in our future work.

1 Introduction

This project seeks to convert research prototypes of spatial interaction components (Janarthanam et al., 2013), into a robust and extendable mobile application. In particular, we aim to produce a technology which is easily portable to new locations and domains. As such, the core technology of this project operates over open data sources which are scraped from the web to create a "CityModel" (Section 2.2). The response time of the system is improved using a client-server architecture, as described in Section 2.1. We also developed a modular, multi-threaded interaction manager where individual modules are transferrable to new domains (Section 2.3). In contrast to existing mobile applications, our Android agent is able to simultaneously interleave multiple tasks, e.g. navigation and tourist information. The system can also take initiative and push relevant information to the user, based on task priority and user interests (Section 3). In this demo, we present a system for tourist information within the city of Edinburgh. The example dialogue in Table 1 illustrates the system's capabilities on navigation, search, weather, events and information push on Points of Interests (PoI).

User:	Take me to the National Museum.
System:	Ok. Here is a route on the map. [*displays map*]
User:	Thanks. Will it rain today?
System:	Showers are forecasted throughout the day.
User:	Oh no. Is there a cafe nearby?
System:	There are several cafes close to you. The closest one is the Elephant House. [*displays locations on the map*]
User:	Is it any good?
System:	It has a 4-star rating on Foursquare.
System:	It also features a literary pub quiz at 6pm today.
User:	[*continues walking*]
System:	On your right you can see a statue of a dog known as Greyfriars Bobby. Do you want me to tell you more?

Table 1: An example SpeechCity interaction.

2 Architecture

The architecture of the system is shown in Figure 1. The application adopts a client-server architecture. The Android mobile app serves as a client and consists of the interface and dialogue system. The server consists of the CityModel.

2.1 Android Interface

The Android interface module manages the GUI display and user input, and serves as the interface between the user and the dialogue system. It accesses the onboard sensors to determine user's positioning and pace and uses Google ASR API for recognising user speech. Th information are sent to the dialogue system as user inputs. It receives dialogue system outputs such as system utterances and information to be displayed on the screen as well as on the map. The interface uses onboard Google TTS for speech synthesis. It uses OpenStreetMaps API to display the map with user location and other information such as location of

restaurants, pubs, etc that the user requested. In addition to spoken interaction, the system also features multi-modal interaction for the user to query the system from a list displayed or by clicking on the map, see Figure 2.

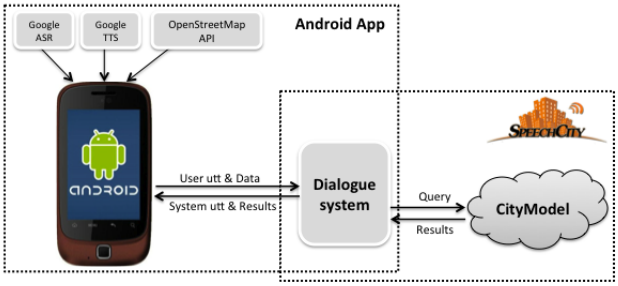


Figure 1: Architecture of SpeechCity mobile application.

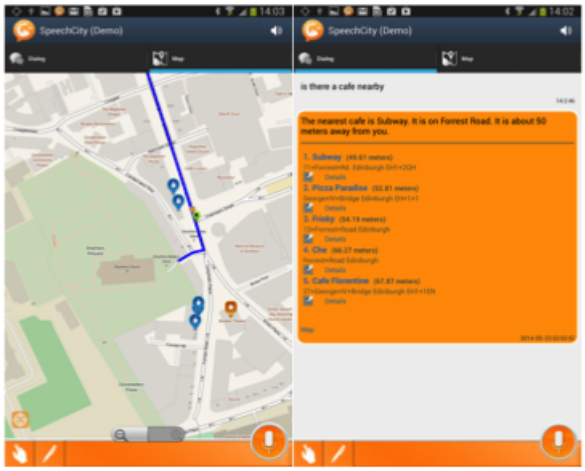


Figure 2: Multimodal SpeechCity interface.

2.2 City Modelling using Open Spatial Data

The CityModel is a spatial semantic database containing information about several hundred thousand Points of Interest in the city of Edinburgh. We harvest these entities from Open Street Maps, which are then annotated with data from social networks (FourSquare) and Wikipedia. These data include location, use class, name, street address, user ratings, number of check-ins, URL, etc. For disambiguating entities across data sources we use several factors including Levenshtein edit distance between names, geographical coordinates, address, phone number and postcode.

2.3 Multi-tasking dialogue system

The Dialogue System, consisting of an utterance parser, an interaction manager, and an utterance generator. Parsing and generation of utterances are done using hand-coded rules.

The Interaction Manager (IM) is the central component of this architecture, which provides the user with navigational instructions, pushes PoI information and manages QA questions. It handles several tasks like navigation, guided tours, weather information, amenity search and PoI information.

The IM receives the user's input in the form of a dialogue act (from the rule-based parser), the user's location (latitude and longitude) and pace rate. The location coordinates of the user are sent to the IM every 2 seconds. In addition, the IM has to deal with incoming requests and responses from the user's spoken inputs. With the possibility of system utterances being generated at a frequency of one every two seconds, there is a need for an efficient mechanism to manage the conversation and reduce the risk of overloading the user with information. In order to manage multiple conversational threads we implemented techniques such as multi-threading, prioritised queue management, and queue revision (Janarthanam and Lemon, 2014). Different dialogue threads are visually represented as separate display cards.

3 Tasks and Features

The system handles a variety of tourist information tasks such as searching for amenities (restaurants, pubs, etc), finding attractions to visit, getting directions, events, weather information, and historical information about monuments, etc. The system pushes information and recommendations according to the user's preferences and interests. For example, it might notify the user of nearby historical sights. We currently obtain user model information from the initial registration phase. In future work, we will acquire such information from the user's utterances and social network and continuously infer and update the model based on previous interactions.

4 Future Work

In future work, we aim to move away from template-based Natural Language Generation towards a domain-general framework using machine learning for generating robust instructions under uncertainty (Rieser and Lemon, 2011; Lemon et al., 2010).

Acknowledgments

The research leading to this work has received funding from the Engineering and Physical Science Research council, UK (EPSRC) under an Impact Acceleration Grant and the EPSRC project no. EP/L026775/1. For further information see `http://speechcity.com/`

References

Srinivasan Janarthanam and Oliver Lemon. 2014. Multi-threaded interaction management for dynamic spatial applications. In *Proceedings of Dialogue in Motion workshop, EACL, 2014*.

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 *Proceedings of the SIGDIAL 2013 Conference*, pages 151–153, Metz, France, August. Association for Computational Linguistics.

Oliver Lemon, Srini Janarthanam, and Verena Rieser. 2010. Generation under uncertainty. In *Proceedings of INLG / Generation Challenges*.

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

Clarification Requests at the Level of Uptake

Julian J. Schlöder and **Raquel Fernández**
Institute for Logic, Language & Computation
University of Amsterdam
julian.schloeder@gmail.com, raquel.fernandez@uva.nl

Abstract

In cooperative dialogue, participants are expected to jointly *take up* each other's moves. The process leading up to uptake can be aided by repair mechanisms, in particular *clarification requests*. We discuss how clarification requests occur after mutual understanding, but before full uptake, and relate them to preparatory conditions of conversational projects.

1 Introduction

Dialogue is frequently viewed as an inherently cooperative activity where interlocutors do not merely exchange singular moves, but actively collaborate in a form of *joint action*. For each utterance put forward in a cooperative dialogue, this process fully succeeds when the addressee *takes up* (her construal of) the speaker's intended act, in which case they are jointly committed to a *joint project* (Clark, 1996). We follow Clark in treating every speech event as a joint project proposal, *e.g.*, an assertion projects adopting its content as mutual belief, and a question projects an answer.

Most work on the process of grounding and clarification has focused on coordination at the levels of perception and understanding (Traum, 1994; Gabsdil, 2003; Ginzburg and Cooper, 2004; Purver, 2004; Schlangen, 2004; Ginzburg, 2012). However, Schlangen (2004) proposes a classification scheme for clarification requests (CRs) that, amongst other dimensions, distinguishes four problem sources for CRs corresponding to the four levels of communication proposed by Clark (1996) and Allwood (1995). In addition, Benotti (2009) is concerned with CRs related to planning which we would (partly) attribute to the fourth level in such a hierarchy (uptake). The excerpt in (1), from the British National Corpus (Burnard, 2000), shows an example of what we consider to be an uptake-level clarification request:

(1) A: I know Vic has cream in his [food] and
B: **How do you know?**
A: Well it said so on the menu, that's why.

2 Types of Uptake CRs

Rodríguez and Schlangen (2004) put forward an annotation scheme based on Schlangen's classification and use it to annotate a portion of the Bielefeld corpus of task-oriented dialogue, where an instruction giver (I) guides an instruction follower (K) through the construction of a paper airplane. They define level 4 CRs as being related to "recognizing or evaluating speaker intention". Examples (2) and (3) below have been classified as level 4 by Rodríguez and Schlangen (2004).[1]

(2) I: you have to put these in between there
K: **in between how?**
I: in between the well you have the wings on top

(3) K: for me that is in fact below this [...]
I: **why below?**
K: yes, it belongs there, all okay.

In (2), K is requesting additional information on what he is to do, but seems generally willing to do what is asked of him. On the other hand, the CR in (3) seems to indicate a general reluctance on I's side to take up K's proposition: K indicates that something is on the wrong side of the plane, but instead of agreeing to this, I questions the reasons K might have for stating this.

Rieser and Moore (2005) annotate the CRs in the Carnegie Mellon Communicator Corpus (Bennett and Rudnicky, 2002) using a refined version of the annotation scheme by Rodríguez and Schlangen (2004); they see problem sources at level 4 not only in intention evaluation, but also in what they call (contradicting) belief (4) and ambiguity refinement (5).

(4) Agent: You need a visa.
Cust: **I do need one?**
Agent: Yes you do.

(5) Agent: [...] that is fifty one dollars.
Cust: **Per day?**
Agent: Per day um mm.

[1] We thank the authors for granting us access to their annotated corpus. The excerpts have been translated to English for the purposes of this abstract by a native speaker of German.

3 Clarification Potential of Uptake

The examples in the previous section show that uptake can fail for different reasons. This points towards different communicative problems arising on that level which may need to be explicitly dealt with; this is what Ginzburg (2012) has described as *clarification potential*. We propose that this potential stems from the failure of some underlying preconditions. We first describe these conditions and then present examples[2] on how they are reflected in CRs on uptake level; also see table 1.

We propose the following conditions, inspired by Clark's (1996, p. 203) discussion of joint purposes, which we take to be common to any project proposal:[3]

- The speaker has sufficient reason to take part in the project (speaker reason),
- the addressee does not have reason not to take part in the project (addressee reason), and
- both speaker and addressee have the requisite knowledge and ability to perform the project (speaker knowledge and addressee knowledge).

The asymmetry in the speaker reason and addressee reason conditions stems from the assumption that addressees are generally cooperative. So they would only refuse collaboration on a joint project if they have grounds *not to*. This effect can be observed in example (6). If the proposed project is an assertion, the typical reason for the failure of the addressee reason condition is that the addressee believes something contrary to the proposal as described by Rieser and Moore (2005).

The exchange in (6) is an example of a CR that is in turn countered by another CR on uptake level:

(6) A: Oh, you can pop in and get your fishing magazines while you're down here
B: **Why?**
A: **Well why not?**

Participant B does not take up A's joint project proposal but instead requests clarification towards the preparatory condition *addressee reason*, asking for a reason to take up the project (*'Why should I do that?'*). This request can be seen as non-cooperative if B is indifferent towards the proposal, as addressees are expected to only clarify this condition if they actually have adversarial motivations. Accordingly, in her response, A does not supply a reason, *i.e.*, does not take up B's CR. Instead, A inquires towards *speaker reason*: what grounds B has for requesting clarification instead of taking up (*i.e.*, what that adversarial reason might be).

CR Type		Example from BNC
Knowledge	Speaker	(1) *How do you know?*
	Addressee	(7) *How [can I tell]?*
Reason	Speaker	(6) *Why not?*
	Addressee	(6) *Why [do this]?*

Table 1: General types of uptake-level CRs with BNC examples; the addressee is the CR initiator.

In example (7), A provides an explanation to an earlier question, *'you can tell'*, but B is unwilling to take this up, and asks a CR towards knowledge.

(7) A: Oh you can tell ⟨pause⟩
B: **How?**
A: against the light

In this case, the surface form *'How?'* is elliptical and could either be towards speaker knowledge (*'How can you tell?'*), addressee knowledge (*'How could I tell?'*), or simply be underspecified (*'How can one tell?'*).

In (8) we have an example where the preparatory condition *addressee ability* truly fails, and the interlocutors collaborate to uncover this.

(8) A: Mummy says you gotta come to her house and pass the things ⟨laugh⟩.
B: No.
A: **No? Why not?**
B: I can't cos I can't open the door.
A: That's alright.

In response to A's CR, B argues that his ability condition fails, so the project cannot be executed.

4 Conclusion

We have surveyed the current work on clarification requests on uptake level, and explained them in terms of general preconditions that apply to both speakers willingly taking up a joint project proposal. We have presented further examples on how these preconditions occur and interact in dialogue. These considerations are part of our investigation into the notion *uptake*; our immediate next goal is a systematic corpus study of these CRs.

[2] All examples are from the BNC (Burnard, 2000) and retrieved with SCoRE (Purver, 2001)

[3] We do not claim that this exhausts possible preconditions; in particular, specific speech events are expected to have more particular conditions.

References

Jens Allwood. 1995. An activity based approach to pragmatics. *Gothenburg papers in theoretical linguistics*, (76):1–38.

Christina L. Bennett and Alexander I. Rudnicky. 2002. The carnegie mellon communicator corpus. In *Proceedings of the International Conference of Spoken Language Processing (ICSLP02)*.

Luciana Benotti. 2009. Clarification potential of instructions. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 196–205.

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

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

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35, Stanford, CA.

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. 2012. *The interactive stance*. Oxford University Press.

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

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, University of London.

Verena Rieser and Johanna D. Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 239–246. Association for Computational Linguistics.

Kepa Joseba Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue.

David Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

Tailoring Object Orientation Descriptions to the Dialogue Context

Gesa Schole
RTG 1808 Ambiguity
University of Tuebingen
gesa.schole@uni-tuebingen.de

Thora Tenbrink
School of Linguistics and English Language
Bangor University
t.tenbrink@bangor.ac.uk

Kenny R. Coventry
Psychology Department
University of East Anglia
k.coventry@uea.ac.uk

Elena Andonova
Cognitive Sciences Department
New Bulgarian University
eandonova@nbu.bg

1 Introduction

Research on the verbal description of object placement has primarily focused on *where* objects are (e.g., Plumert et al., 1995), disregarding how they are *oriented*. In the present study, we describe when dialogue partners *exchange* object orientation information in a referential communication task, or rather rely on *inferences*.

In dialogue, speaker and addressee try to keep an idea about common ground to guarantee understanding (Clark, 1996). According to the 'Principle of least collaborative effort', both dialogue partners try to minimize the conversational effort for themselves and for their partner at the same time (Clark and Wilkes-Gibbs, 1986), and expect their dialogue partner to draw inferences from common ground (Spenader, 2002). Levinson (2000, 32) suggests that inferences from cultural knowledge are licensed by the *I-heuristic*, namely: "What is simply described is stereotypically exemplified". A contribution should therefore not be more informative than is required in a particular conversational situation (see also Grice, 1975), because minimal descriptions (of object orientation) may already license stereotypical interpretations of the situation.

2 Empirical Study

Our empirical study (first reported in Tenbrink et al., 2008) was a referential communication task. In the study, the *director* described for the *matcher* how to furnish an empty dolls' house. The director's dolls' house was pre-furnished either in a functional array (*f*), in which the rooms could be identified as kitchen, living-room, bedroom, and bathroom, or in a non-functional array (*nf*), in which the rooms could not be associated with a specific function (see Figure 1). The participants were separated by a screen, so that they could not see each other or the interior of their partner's dolls' house. The present analysis comprises the data of sixteen randomly selected dyads (eight dyads per condition).

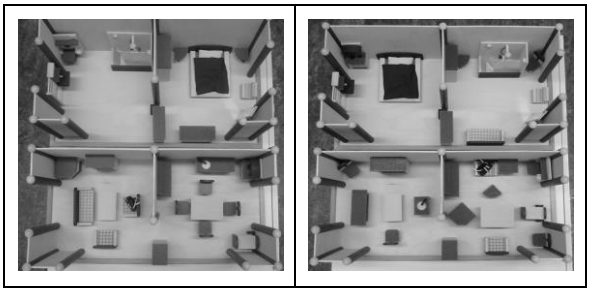


Figure 1. The model houses (left: *f*, right: *nf*).

2.1 Coding of Orientation Errors

For the present analysis, orienting objects correctly provided a measure for evaluating communication success. Objects were considered as oriented incorrectly when their orientation differed from the model by more than 45°. Error scores were coded by two independent raters who agreed in 96.19% of cases. Coding disagreement was resolved by a third coder.

2.2 Annotating Orientation Information

Referring to an object's orientation may involve its geometric properties, such as axes, which are projected onto objects analogous to the human body's axes (Landau and Jackendoff, 1993). Descriptions were considered *complete* when they included explicit references to one of the locatum's (directed) axes or features, a spatial term, a reference object to describe orientation if required by the spatial term (e.g., *towards the bed*), and (if required) reference to diagonal orientation. Based on this annotation, orientation descriptions were

evaluated as being *complete*, *incomplete*, *depending*, or *missing* for each object individually.

Example (1) shows a *complete* description of a shelf, and (2) exemplifies an *incomplete* description. A few descriptions such as (3) *depended* on the orientation of an object described earlier; this could lead to placement errors if the previous object was placed incorrectly.

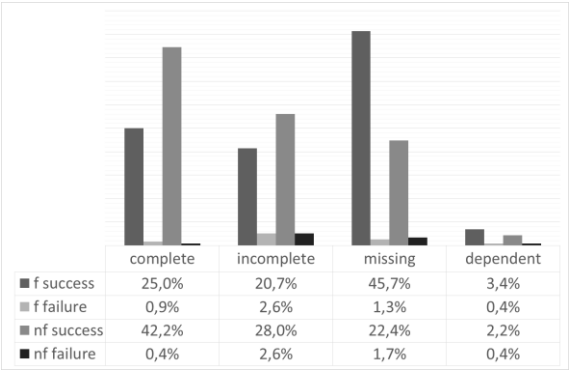
(1) *director*: ein großes Regal (…) mit den blauen Türen zum Bett rüber
[a big shelf (…) with the blue doors towards the bed]

(2) *director*: in die Ecke ist schraeg ein Stuhl reingestellt
[there is a chair placed diagonally into the corner]

(3) *director*: äh die Toilette is äh parallel zur Dusche praktisch an die Hinterwand gestellt.
[uh the toilet is uh parallel to the shower practically placed at the back wall.]

3 Results

In each of the conditions (*f* and *nf*), there were only 12 orientation errors out of 232 objects to be placed. Figure 2 shows our four categories for orientation information (*complete*, *incomplete*, *missing*, *dependent*) according to condition (*f* vs. *nf*), and further distinguishes between success and failure to orient the object in focus correctly. The results presented in the following focus on the discrepancy between the two conditions regarding orientation success based on *complete* orientation information and *missing* references.



	complete	incomplete	missing	dependent
f success	25,0%	20,7%	45,7%	3,4%
f failure	0,9%	2,6%	1,3%	0,4%
nf success	42,2%	28,0%	22,4%	2,2%
nf failure	0,4%	2,6%	1,7%	0,4%

Figure 2. Extent of orientation information (per object).

While *complete* orientation information was given for 58 objects (25.0%) that were then successfully placed in the functional condition, this applied to 98 objects (42.2%) in the non-functional condition. Conversely, orientation information was *missing* for 106 objects (45.7%) that were successfully placed in the functional condition, but for only 52 objects (22.4%) in the non-functional condition. The overall pattern was highly significant ($\chi^2 = 29.50$, df = 3, $p<.001$).

4 Discussion

Our data demonstrate that both director and matcher were sensitive to the availability of cultural knowledge of functional relations between objects. Directors and matchers adjusted their descriptions of orientation information to context-specific conditions, and matchers regularly made correct inferences from missing orientation information, using their cultural knowledge to fill in the gaps. While dialogue partners tended to explicitly negotiate orientation information for atypical spatial arrangements, in the case of typical object arrangements they relied on inferences drawn from cultural knowledge far more often. Based on this adaptation to the availability of cultural knowledge, errors occurred seldom and to the same extent in both conditions – irrespective of the typicality of the spatial situation. Clearly, functional arrangements supported and simplified communication, adding to previous findings on effects of functional relationships (e.g., Coventry and Garrod, 2004).

These findings comply with Clark and Brennan's (1991) suggestion that information is communicated when perceived as necessary. They also provide an exemplification of the I-heuristics (Levinson, 2000). In our data, director and matcher mostly relied on the I-heuristic for individual objects when the spatial array was stereotypical. However, they tended to rely on verbal information exchange when the spatial arrangement was atypical. This was the case even though the objects in our study were, in fact, all set in a typical orientation; no object was oriented towards the wall or put upside down. In this way, use of the I-heuristic appeared to be mediated by the typicality of the object arrangement. With a non-typical arrangement, speakers apparently felt that object orientation could not be left out, leading to less 'simple' descriptions (in Levinson's terms) and, accordingly, less stereotypical interpretations. Still, even in atypical situations, matchers were able to make appropriate inferences. Thus, common ground plays a crucial role for inferring or interpreting information about object orientation in all situations.

In future research we aim to investigate the strategies of the dynamic dialogue processes in this regard in more detail, towards further insight into how joint dialogic effort ties in with conversational inference processes.

References

Herbert H. Clark. 1996. Using Language. Cambridge: Cambridge University Press.

Herbert H. Clark and Susan A. Brennan. 1991. Grounding in Communication. In: Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (Eds.). Perspectives on Socially Shared Cognition. Washington: APA Books. 127-149.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring As a Collaborative Process. *Cognition*, 22(1):1-39.

Kenny R. Coventry and Simon C. Garrod. 2004. Saying, Seeing, and Acting. The Psychological Semantics of Spatial Prepositions. Hove, East Sussex, New York: Psychology Press.

H. Paul Grice. 1975. Logic and conversation. In: Peter Cole, Jerry Morgan (Eds.). Syntax and Semantics. New York, San Francisco, London: Academic Press. 41-58.

Barbara Landau and Ray Jackendoff. 1993. "What" and "Where" in Spatial Language and Spatial Cognition. *Behavioural and Brain Sciences*, 16:217-265.

Stephen C. Levinson. 2000. Presumptive Meanings: The Theory of Generalized Conversational Implicature. Cambridge, Mass: MIT Press.

Jodie M. Plumert, Christopher Carswell, Kathy de Vet, and Damien Ihrig. 1995. The Content and Organization of Communication About Object Locations. *Journal of Memory and Language*, 34:477-498.

Jennifer Spenader. 2002. Presupposed Propositions in a Corpus of Dialogue. In: Kees van Deemter and Rodger Kibble (Eds.). Information Sharing: Reference and Presupposition in Language Generation and Interpretation. Stanford: CSLI Publications.

Thora Tenbrink, Elena Andonova, and Kenny R. Coventry. 2008. Negotiating Spatial Relationships in Dialogue: The Role of the Addressee. In: Jonathan Ginzburg, Pat Healey, and Yo Sato (Eds.). Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue. LONDIAL. King's College, 2 - 4 June, 201-208.

Perception Based Misunderstandings in Human-Computer Dialogues

Niels Schütte
Dublin Institute of Technology
niels.schuette@student.dit.ie

John Kelleher
Dublin Institute of Technology
john.d.kelleher@dit.ie

Brian Mac Namee
Dublin Institute of Technology
brian.macnamee@dit.ie

Abstract

In a situated dialogue, misunderstandings may arise if the participants perceive or interpret the environment in different ways. In human-computer dialogue this may be due the sensor errors. We present an experiment system and a series of experiments in which we investigate this problem.

1 Introduction

Computer systems that engage in natural language dialogue with human users are known as **dialogue systems**. A dialogue system that operates in a spatial environment, a **situated dialogue system**, needs to have information about the spatio-temporal context. This can be achieved through perception of the environment. Perception, e.g. computer vision, always has the potential of producing errors, e.g. by failing to notice an object or by misrecognizing an object as a different type of object. We are interested in the effect that such perception-based errors have on human-computer dialogue. If the human user and the system have a shared view of the environment, false perception by the system will lead to a divergence between the user's understanding of the environment and the system's understanding and this in turn leads to problems in the interaction between the system and the user. For example, if the user asks a robot to pick up an apple, and the robot has mistaken a pear for an apple, it may instead pick up the pear. Misunderstandings of this kind also occur in human-human interaction and human speakers are able to establish and recover a shared understanding or common ground (Clark and Schaefer, 1989). Misunderstandings in human-computer dialogue due to misunderstandings because of problems in natural language understanding and speech recognition have been also been investigated and addressed (e.g. (Shin et al., 2002; López-Cózar et al., 2010)).

In an earlier work we investigated the problem of perception based misunderstandings in a corpus of data from human-human interaction (Schütte et al., 2012). In this paper we report on a work in progress in which we investigate the effect of sensor errors on human-computer dialogue using a dialogue system. Participants interact with a simulated robot through a text based dialogue interface in order to re-arrange objects in a virtual world.

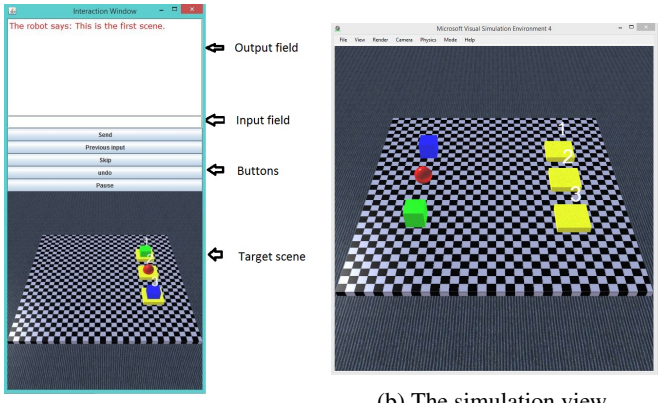
Participants are presented with a number of **scenes**. In each scene the participants are asked to instruct the robot arrange the objects present in the world into a given **target configuration**. Participants were given the option to abandon a scene if they felt they would not be able to complete it.

We perform a series of experiments that focus on three issues: (a) establishing a **baseline** for the difficulty of the interaction task (b) establishing the **impact of perception errors** on the baseline task performance and (c) establishing the **usefulness of different approaches to resolve the misunderstandings**.

2 Experiment System

The experiment system consists of a dialogue system and a robot simulation environment. The dialogue system was implemented for this experiment and is focused on covering a wide range of spatial instructions. The robot is a highly simplified abstraction of a manipulator arm that can pick up objects and move them to specified locations. It is not rendered in the simulation.

Figure 1 shows the user interface presented to the participants. The left side shows the text based dialogue interface window. The image in the lower part of the window shows the scene the participant was asked to create. The right hand side of the figure shows the participants' view of the simulated world. In summary, the participants' task was to



(a) The dialogue interface.

(b) The simulation view.

Figure 1: The user interface.

interact with the system to change the scene in the window in Figure 1b into the scene displayed at the bottom of the window in Figure 1a.

The robot's perception of the world is provided through an abstract simulated vision system. By manipulating the vision system, targeted errors can be introduced into the system's perception. For example, it can be specified that the system mistakes the colour of a certain object for a different colour. If the participant now uses the colour to describe the object, the system will not be able to resolve the reference correctly. It should be noted that with this experiment we do not aim to produce a novel dialogue system or to provide an accurate simulation of computer vision, but to examine the performance of the the given system under different conditions of perceptual problems.

During each interaction, the contributions by the participant and the system are logged and annotated with their semantic interpretations. Parameters related to the dialogue such as task completion rate, number of actions, number of errors and time taken for each action are recorded. They serve as the basis of our comparison of the task difficulty of the different experiment conditions.

We are currently performing a series of three experiments with this experiment setup. **Experiment 1** serves to establish a baseline difficulty. It uses a series of scenes that were manually designed to encourage specific expressions. In **Experiment 2** errors are introduced into the robot's perception. This experiment serves to establish the impact of the errors on the interaction. Errors were manually designed for each scene to produce specific problem situations. In **Experiment 3** we evaluate different approaches towards solving the perception based misunderstandings by communicating the system's understanding of the scene to the user. The experiment uses the same scenes and errors as the second experiment. Participants are split into two groups. The first group is given the option of asking the system to describe verbally what it perceives. The second group is given the option of asking the system to visually communicate its understanding through the use of markup on the screen. Thereby both groups are given access to the system's understanding of the scene, but through different modalities.

3 Current State

We have currently finished the first two experiments and are evaluating the results. A first preliminary analysis and a more detailed description of the experiment will be available in (Schütte et al., 2014). The third experiment is currently commencing. A comparison of the results from the first and the second experiment indicates that the introduction of perception errors increased the difficulty of the task. Participants were much more likely to abandon scenes containing errors than scenes not containing errors. They also needed more actions to complete scenes with errors than scenes without errors, and often used more time doing so.

4 Future Work

After the completion of the third experiment, we are going to compare the results between the different experiments. We are planning to investigate the strategies used by the participants when they encountered problems in the dialogue and relate them to our work in human-human interaction.

References

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, pages 259–294.

Changsong Liu, Rui Fang, and Joyce Y. Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 140–149. Association for Computational Linguistics.

Ramón López-Cózar, Zoraida Callejas, and David Griol. 2010. Using knowledge of misunderstandings to increase the robustness of spoken dialogue systems. *Knowledge-Based Systems*, 23(5):471–485, July.

Niels Schütte, John Kelleher, and Brian Mac Namee. 2012. A corpus based dialogue model for grounding in situated dialogue. In *Proceedings of the 1st Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Language, Motor Control and Vision (MLIS-2012).*, Montpellier, France, August.

Niels Schütte, John Kelleher, and Brian Mac Namee. 2014. The effect of sensor errors in situated human-computer dialogue. In *Proceedings of the The 3rd Workshop on Vision and Language (VL'14).*, page to appear, Dublin, Ireland, August.

Jongho Shin, Shrikanth S. Narayanan, Laurie Gerber, Abe Kazemzadeh, Dani Byrd, and others. 2002. Analysis of user behavior under error conditions in spoken dialogs. In *INTERSPEECH*.

Sample Efficient Learning of Strategic Dialogue Policies

Wenshuo Tang, Zhuoran Wang, Verena Rieser, and Oliver Lemon
Interaction Lab
Heriot-Watt University
Edinburgh, UK
Email: wt92@hw.ac.uk Website: www.macs.hw.ac.uk/InteractionLab

Abstract

This work aims to learn strategic dialogue policies which estimate the (hidden) state of the opponent, using extensions of Partially Observable Markov Decision Processes. As a first step towards this goal, we present results of batch Reinforcement Learning (LSTD), which needs only 20% of the training data needed by SARSA(λ). This result now puts us in the position to tackle more computationally intensive partially observable environments .

1 Introduction

Strategic dialogue behaviour includes cooperative as well as non-cooperative actions and the ability to choose amongst these actions dependent on the current context, which includes your long-term goal and current state, as well as the goal and state of your interaction partner (also known as the "opponent"). In this research we investigate opponent modelling for optimising strategic dialogue using models based on Partially Observable Markov Decision Processes (POMDPs), following an initial proposal by (Rieser et al., 2012). Recent work has shown that the ability to reason about each other's beliefs (in terms of states and goals) using Decentralised POMDPs, enables agents to evolve cooperative behaviour (Vogel et al., 2013a), resolve implicatures (Vogel et al., 2013b), and reason about acceptable actions towards a human collaborator (Kamar et al., 2013). We hypothesise that this ability will also allow us to learn strategic dialogue policies which reason about the opponent's state. However, these types of extended POMDP models (and POMDPs in general) are intractable for more complex domains and approximate models are used in practise. Furthermore, they require efficient training algorithms to solve the underlying POMDP. Previous work on non-collaborative dialogue has found that it takes about **100k** of training games to learn a policy that can beat a rule-based opponent in a fully supervised MDP setting with a state space size of 16k (Efstathiou and Lemon, 2014). This previous work has used a online Reinforcement Learning algorithm called SARSA(λ). Current research on POMDPs for statistical dialogue management investigates more sample efficient algorithms such as GPTD (Gasic and Young, 2014), KTD (Daubigney et al., 2012) or LSPI (Pietquin et al., 2011).

In the following, we explore a combination of function approximation methods and offline learning, using batch Least-Squares Temporal Difference (LSTD) approximation. We evaluate this approach against previous work by (Efstathiou and Lemon, 2014) using the same experimental setup within a strategic trading game.

2 The Testbed Trading Game

Taikun is a 2-player, sequential, non-zero-sum game with imperfect information designed to investigate non-cooperative dialogue in a controlled environment (Efstathiou and Lemon, 2014). The goal of the game is for each participant to collect resources (Rock, Wheat and Sheep) via trading or by random game update. In the trading phase a player proposes a 1-for-1 trade of resources and the other player accepts or rejects the proposed trade. In the game update phase the environment randomly modifies the resources of each player by adding two or subtracting one. This information is hidden to the other player. The setup also includes a challenging rule-based adversary which wins 66% games against a random policy. Further details on the adversary's policy can be found in (Efstathiou and Lemon, 2014). The goal state of the Learning Agent (LA) and adversary are predefined and partially overlapping, as shown in Table 1, which motivates trading.

	Wheat	Rock	Sheep
LA	4	5	0
Adv.	4	0	5

Table 1: Goal state for Learning Agent and Adversary

3 Experiment setup and results

We now test different parameterisations of a sample efficient reinforcement learning algorithm called Least-Squares Temporal Difference (LSTD) (Bradtke and Barto, 1996), which is an offline function approximation approach. We evaluate these algorithms using the same setup as (Efstathiou and Lemon, 2014), where we formulate the problem as Markov Decision Process (MDP). The state is represented by the LA's set of resources (only) and the actions are 7 different trading offers (do nothing, trade X resource for Y resource). The long term reward is $+1000$ for winning a game, $+500$ for a draw and -100 for losing. We evaluate the learnt policies on 50k test games. The results are summarised in Table 2.

LA Policy	LA	Adversary	Draws	# games
SARSA(λ)	49.23%	45.62%	5.15%	100k
LSTD	44.5%	51.32%	4.18%	5k
Batch LSTD	46.31%	50.76%	2.93%	17k
Batch LSTD*	48.82%	48.03%	3.05%	20k

Table 2: Winning rates for Learning Agents (LA) trained with different algorithms.

First experiment : LSTD learning agent. For the first experiment we experiment with a 'vanilla' version of LSTD using the same state space factorisation as (Efstathiou and Lemon, 2014). The offline training data is generated by a random policy interacting against the rule-based adversary. In this experiment, the adversary outperforms the LA with 51.32% winning rate. The learning curve shows that the LSTD plateaus after 5k training games. We attribute this early convergence towards a non-optimal policy to the fact that LSTD learns from random data. In other words, since off-line algorithms do not have the capability to explore and exploit, the algorithms does not "see" enough instances of the optimal policy.

Second experiment : Batch learning. In a second experiment we use batch reinforcement learning to enhance exploitation, i.e. interleaving a piece-wise online data collection with offline learning (Lange et al., 2012). That is, it combines the policy-search efficiency of policy iteration with the data efficiency of LSTD. We start from an initial policy (LSTD policy trained on 1k games) interacting with our rule based adversary. We then iterate policy learning and data collection, where we use the latest policy to generate new training data for the next learning phase. The results show that batch LSTD has better sample efficiency and reaches higher performance than vanilla LSTD but still falls behind the adversary. We hypothesise that this is due to the insufficient representation of discriminative state features. For example, a required resource will consistently have a positive contribution to the "trade-in" action regardless of whether its amount already exceeds the goal.

Third experiment : Batch policy learning with non-linear state factorisation. We now experiment with a different state space factorisation to (Efstathiou and Lemon, 2014), where we represent the distance from the goal state. In particular, the state contains 6×3 binary variables, recording each individual resource for each of the 3 types as 0/1 up to a maximum of 5 (which is the max in the goal state). The 6th variable indicates whether the agent holds more than 5 of a given resource. The results for the re-factored Batch LSTD* show that the learned policy now performs equal to the challenging rule-based adversary and reaches a similar performance to SARSA(λ) after only 20k games, rather than 100k games[1].

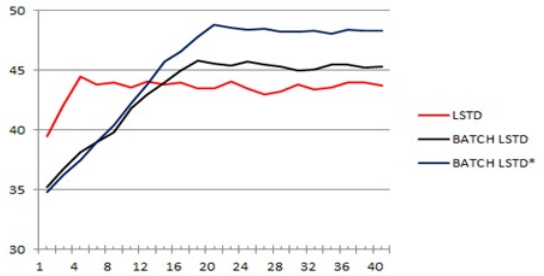


Figure 1: Learning curve: Reward over training data.

4 Discussion and future work

In this paper we have shown that it is possible to learn strategic dialogue policies which can reach a similar performance to a challenging rule-based adversary from a (relatively) small amount of training data (20k games). This now puts us in the position to tackle a more challenging problem where we account for the uncertainty in adversary's state by modelling the problem as a Partially Observable Markov Decision Process.

[1] In future work we will establish statistical significance between winning rates.

Acknowledgments

The research leading to this work has received funding from the European Community's ERC programme under grant agreement no. 269427 (STAC). http://www.irit.fr/STAC/

References

Steven J Bradtke and Andrew G Barto. 1996. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1-3):33–57.

L. Daubigney, M. Geist, and O. Pietquin. 2012. Off-policy learning in large-scale pomdp-based dialogue systems. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, pages 4989 – 4992, Kyoto, Japan. IEEE.

Ioannis Efstathiou and Oliver Lemon. 2014. Learning non-cooperative dialogue behaviours. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 60–68, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

M. Gasic and S. Young. 2014. Gaussian processes for pomdp-based dialogue manager optimization. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(1):28–40, Jan.

Ece Kamar, Ya'akov (Kobi) Gal, and Barbara J. Grosz. 2013. Modeling information exchange opportunities for effective human-computer teamwork. *Artificial Intelligence*, 195(0):528 – 550.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pages 45–73. Springer Berlin Heidelberg.

Olivier Pietquin, Matthieu Geist, Senthilkumar Chandramohan, and Hervé Frezza-Buet. 2011. Sample-efficient batch reinforcement learning for dialogue management optimization. *TSLP*, 7(3):7.

Verena Rieser, Oliver Lemon, and Simon Keizer. 2012. Opponent modelling for optimising strategic dialogue. In *The 16th workshop on the semantics and Pragmatics of Dialogue (SeineDial'12)*.

Adam Vogel, Max Bodoia, Christopher Potts, and Dan Jurafsky. 2013a. Emergence of gricean maxims from multi-agent decision theory. In *Proceedings of NAACL 2013*.

Adam Vogel, Christopher Potts, and Dan Jurafsky. 2013b. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August. Association for Computational Linguistics.

PDRT-SANDBOX: An implementation of Projective Discourse Representation Theory

Noortje J. Venhuizen
University of Groningen
n.j.venhuizen@rug.nl

Harm Brouwer
Saarland University
brouwer@coli.uni-saarland.de

Abstract

We introduce PDRT-SANDBOX, a Haskell library that implements Projective Discourse Representation Theory (PDRT) (Venhuizen et al., 2013), an extension of Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 1993). The implementation includes a translation from PDRT to DRT and first-order logic, composition via different types of merge, and unresolved structures based on Montague Semantics (Muskens, 1996), defined as Haskell functions.

1 Introduction

The semantic property of projection, traditionally associated with presuppositions, has challenged many structure-driven formal semantic analyses. Linguistic content is said to project if it is interpreted outside the scope of an operator that syntactically subordinates it. In semantic formalisms, this behaviour has often been treated as a *deviation* from standard meaning construction, despite the prevalence of expressions exhibiting it (van der Sandt, 1992; Geurts, 1999; Beaver, 2001). By contrast, we have proposed a formalism that *centralizes* the property of projection as a strategy for integrating material into the foregoing context. This formalism is called Projective Discourse Representation Theory (PDRT) (Venhuizen et al., 2013), and is an extension of the widely used framework Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 1993). In PDRT, all linguistic material is associated with a *pointer* to indicate its interpretation site. In this way, an explicit distinction is made between the surface form of an utterance, and its logical interpretation. The formalism can account for various projection phenomena, including presuppositions (Venhuizen et al., 2013) and Potts' (2005) conventional implicatures (Venhuizen et al., 2014), and has already been integrated into the Groningen Meaning Bank (Basile et al., 2012).

Critically, adding projection pointers to all linguistic material affects the formal properties of DRT non-trivially; the occurrence of projected material at the interpretation site results in non-hierarchical variable binding, and violates the traditional DRT notion of context accessibility, thereby compromising the basic construction mechanism. Here, we present an updated construction mechanism as part of a Haskell library called PDRT-SANDBOX that implements PDRT, as well as standard DRT. The implementation incorporates definitions for building and combining structures, translating Projective Discourse Representation Structures (PDRSs) to Discourse Representation Structures (DRSs) and first-order logic (FOL) formulas, and dealing with unresolved structures via lambda abstractions (Muskens, 1996). Moreover, it allows for various input and output representations, and is highly modular, thereby providing a full-fledged toolkit for use in other NLP applications.

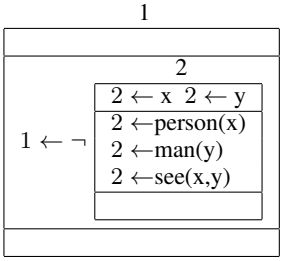
2 Projective Discourse Representation Theory

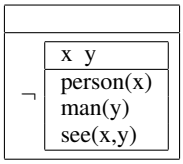
PDRSs carry more information than DRSs; in addition to the structural and referential content of a DRS, a PDRS also makes the information structure of a discourse explicit by keeping linguistic content at its introduction site, and indicating the interpretation site via a projection variable. That is, each PDRS introduces a *label* that can be used as an identifier, and all of its referents and conditions are associated with a *pointer*, which is used to indicate in which context the material is *interpreted* by means of binding it to a context label.

Examples (1) and (2) show two PDRSs and their corresponding DRSs. An important addition to the PDRS definitions described in Venhuizen et al. (2013), is the introduction of *Min-*

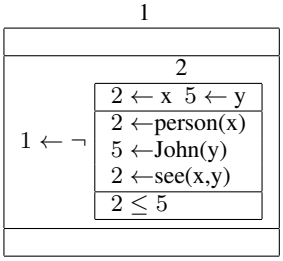
imally Accessible Projection contexts (MAPs) in the footer of each PDRS. These MAPs pose minimal constraints on the accessibility of projection contexts, creating a partial order over PDRS contexts (Reyle, 1993; Reyle, 1995).

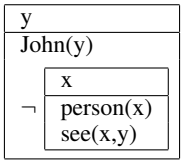
(1) Nobody sees a man.

a.


b.


(2) Nobody sees John.

a.


b.


In the PDRS in (1a), all pointers are bound by the label of the PDRS in which the content is introduced, indicating *asserted material*. As shown in (1b), this representation is identical to the standard DRT representation of this sentence, except for the addition of labels to PDRSs and pointers to all referents and conditions. In (2), on the other hand, the proper name "John" triggers a presupposition about the existence of someone called 'John'. The pointer associated with the referent and condition describing this presupposition indicates *projected material*; it occurs free, as it is not bound by the label of any accessible PDRS. This means that no antecedent has been found yet. In the corresponding DRS in (2b) the presupposition is accomodated at the most global accommodation site. Note that in contrast to the DRT representation, the accommodation site of the presupposition is not determined in the PDRS; (2a) only stipulates that the accommodation site should be accessible from the introduction site of the presupposition. This flexibility of interpretation increases the compositionality of PDRT, since more context may become available later on in which the presupposition becomes bound. In combination with MAPs, this property can also be exploited to account for the projection behaviour of conventional implicatures (Venhuizen et al., 2014).

3 Playing in the PDRT-SANDBOX

We implemented the formal definitions for the construction and manipulation of the structures of PDRT and standard DRT in a Haskell library called PDRT-SANDBOX. For a full description of all definitions, see Venhuizen et al. (in prep). The library provides the following core features:

- **Definitions for building and combining (P)DRSs.** The binding and accessibility definitions in DRT and PDRT are fully worked out, and applied as conditions on combining (*merging*) structures and resolving them. Two different types of merge are defined for PDRT: *projective merge* and *assertive merge* (Venhuizen et al., 2013).
- **Translations.** PDRSs can be translated to DRSs, FOL-formulas, and flat (non-recursive) representations called P-Tables.
- **Lambda abstractions.** Unresolved structures obtain Montague-style representations, following Muskens (1996). The implementation exploits Haskell's lambda-theoretic foundations by formalising unresolved structures as Haskell functions, thereby profiting from all existing associated functionality.
- **Various input and output formats.** As (P)DRS output format, the standard "boxes" representation is available, as well as a linear representation of the boxes, a set-theoretic representation, and the internal syntax for (P)DRSs. The latter two are also recognised as input formats, along with the Prolog syntax from Boxer (Bos, 2003).

4 Conclusion

PDRT-SANDBOX is a full-fledged NLP library for constructing and manipulating the discourse structures from DRT and PDRT, which can be used as part of a larger NLP architecture. One direction would be combining the implementation with a syntactic parser, resulting in a tool-chain similar to the one created by the C&C tools and Boxer (Curran et al., 2007). Furthermore, the representations produced by PDRT-SANDBOX may be applied in a separate model checker, QA system, or any other NLP tool that uses deep semantic representations. PDRT-SANDBOX is freely available (under the Apache License, Version 2.0) at: `http://hbrouwer.github.io/pdrt-sandbox/`

References

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Joost Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3196–3200, Istanbul, Turkey. European Language Resources Association (ELRA).

David I. Beaver. 2001. *Presupposition and Assertion in Dynamic Semantics*. CSLI Publications, Stanford, CA.

Johan Bos. 2003. Implementing the binding and accommodation theory for anaphora resolution and presupposition projection. *Computational Linguistics*, 29(2):179–210.

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.

Bart Geurts. 1999. *Presuppositions and pronouns*. Elsevier.

Hans Kamp and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Hans Kamp. 1981. A theory of truth and semantic representation. In J.A.G. Groenendijk, T.M.V. Janssen, and M.B.J. Stokhof, editors, *Formal Methods in the Study of Language 135*, pages 277–322. Mathematisch Centrum.

Reinhard Muskens. 1996. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy*, 19(2):143–186.

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford University Press, USA.

Uwe Reyle. 1993. Dealing with ambiguities by underspecification. *Journal of Semantics*, 10(2):123–179.

Uwe Reyle. 1995. On reasoning with ambiguities. In *Proceedings of the seventh European chapter of the Association for Computational Linguistics*, pages 1–8. Morgan Kaufmann Publishers Inc.

Rob van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9:333–377.

Noortje J. Venhuizen, Johan Bos, and Harm Brouwer. 2013. Parsimonious semantic representations with projection pointers. In Katrin Erk and Alexande Koller, editors, *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 252–263, Potsdam, Germany, March. Association for Computational Linguistics.

Noortje J. Venhuizen, Johan Bos, Petra Hendriks, and Harm Brouwer. 2014. How and why conventional implicatures project. In *Proceedings of the 24rd Semantics and Linguistic Theory Conference (SALT 24)*, New York, May 30 – June 1. New York University.

Noortje J. Venhuizen, Johan Bos, Petra Hendriks, and Harm Brouwer. in prep. Harnessing projection: implementing Projective Discourse Representation Theory. Manuscript in preparation.

Detecting Deception in Non-Cooperative Dialogue: A Smarter Adversary Cannot be Fooled That Easily

Aimilios Vourliotakis, Ioannis Efstathiou, and Verena Rieser
Interaction Lab
Heriot-Watt University
Edinburgh, UK
www.macs.hw.ac.uk/InteractionLab

Abstract

Recent work has learned non-cooperative dialogue behaviour within a stochastic trading game, including dialogue moves such as bluffing and lying. Here, we introduce an adversary which can detect deception based on logical contradictions between dialogue moves. Being caught in deception, the adversary will penalise this behaviour by either refusing to trade or declaring victory. We compare our results to a learning agent trained with a gullible adversary and show that a more realistic adversary decreases the chances of winning by over 20%, if the penalty for cheating is to lose the game. In future work we will re-train the learning agent within this more challenging environment.

1 Introduction

Deception in artificial agents has been identified as important in variety of application areas, including education, military operations, video games and healthcare (Traum, 2008; Shim and Arkin, 2013). Recently, dialogue policies have been developed which can execute such non-collaborative behaviour using Reinforcement Learning (Georgila and Traum, 2011; Efstathiou and Lemon, 2014). In this research, we extend previous work by (Efstathiou and Lemon, 2014) by evaluating the learnt policy against an adversary which is able to detect deception based on logical inconsistencies between dialogue moves. In contrast, (Efstathiou and Lemon, 2014) have used a simple frequency-based approach, where the likelihood of detection linearly increases the more the agent lies or bluffs.

In the following, we first summarise the learning framework within a stochastic trading game (Section 2). We then describe three models of detecting deception (Section 3). In Section 4 we present some preliminary results, testing the trained learning agent from (Efstathiou and Lemon, 2014) against our extended adversary models.

2 Learning Non-Cooperative Behaviour in Taikun

Taikun is a 2-player, sequential, non-zero-sum game with imperfect information designed to investigate non-cooperative dialogue in a controlled settings environment (Efstathiou and Lemon, 2014). The goal of the game is for each participant to collect resources (Rock, Wheat and Sheep) via trading or by random game update. In the trading phase a player proposes a 1-for-1 trade of resources and the other player accepts or rejects the proposed trade. In the game update phase the game randomly modifies the resources each player has by adding two or subtracting one of them.

In (Efstathiou and Lemon, 2014) a Learning Agent (LA) is modelled as a Markov Decision Process (MDP) and is trained using SARSA(λ) against a rule-based adversary. In order to introduce deception, the LA was supplemented with additional Manipulation Actions (MAs) in the form of "I really need X", where X is a type of resource. The adversary will then adapt its strategy to not engage in or propose trades where the LA would receive this resource. The LA uses these MAs against the "gullible" adversary in order to mislead him into trading resources he actually needed (**Baseline Scenario**). An advanced scenario introduces a risk of deception detection, where the likelihood of discovery by the adversary is increased after each MA (**Frequency-based Approach**).

3 Detecting Deception

Here we detect deception based on a model of semantic inconsistencies (e.g. contradictions) between dialogue moves. The following examples show how deception could be detected:

Scenario	LA wins		ADV wins		Draws	
Baseline (no detection)	59.170		39.755		1.075	
Detection by:	Refusal to trade	Automatic win	Refusal to trade	Automatic win	Refusal to trade	Automatic win
Case1: Plain Lies	55.725	39.996	42.295	58.895	1.980	1.110
Case1+2: Naive Turn	54.035	35.950	43.920	62.945	2.045	1.105
Case1+3: Probabilistic Turn	54.275	36.985	43.810	62.025	1.915	0.990
Frequency-based	50.86	49.7	46.33	46.225	2.81	4.075

Table 1: Winning rates in % for different adversary models

(1) a. LA: I really need Wheat. (**MA**)
b. ADV: I give you Rock and I need Wheat.
c. LA: Ok! (**Contradiction**)
d. (*Game update*)
e. LA: I give you Wheat and I need Sheep. (**Contradiction**)

Note that in real world face-to-face spoken interaction, deception can also be detected from multimodal cues (Fitzpatrick et al., 2012). In our simulations we consider the following cases:

Case 1: Lies in the same trading-phase (Plain Lies). In Example 1 (a) the LA falsely declares that he needs wheat, while in the next dialogue turn it clearly contradicts itself by giving this resource away, see Example 1 (b).

Case 1+2: Lies in consecutive trading-phase (Naive Turn-based Approach). In addition to Case 1, we consider logical inconsistencies which occur between an MA and a subsequent LA action in the next trading phase, see Example (1e). In this case, we ignore the game update in (1d).

Case 1+3: Likelihood of consecutive lies (Probabilistic Turn-based Approach). This case now accounts for the game update, where the LA randomly receives/ loses resources and thus the probability the MA is still valid decreases by 1/3.

Once a MA is discovered, the lie can be penalised in two different ways, following (Efstathiou and Lemon, 2014):
Refusal to trade: After detecting a MA, the adversary will refuse to further trade with the LA.
Automatic win: After detecting a MA, the adversary will win automatically.

4 Results

We now test the trained learning agent ('Baseline') from (Efstathiou and Lemon, 2014) against our extended adversary models. The results in Table 1 show:

- As expected, the LA trained for a gullible adversary performs worse with adversaries which can detect deception.
- Within our three different cases, detecting plain lies within the same turn has the most effect. There is a negligible difference between detecting lies in consecutive turns between the naive approach (Case 2) and the approach which takes environmental uncertainty into account (Case 3).
- Surprisingly, the adversaries which can detect MAs based on logical contradictions perform worse than the frequency-based adversary. However, note that in this case (Efstathiou and Lemon, 2014) actually re-trained the LA and thus the LA had the chance to adapt to this more challenging scenario, so there is no direct comparison. This difference is highlighted by greying out this result in Table 1.
- Finally, when comparing the effect of penalties, we find that refusal to trade has less impact than automatic win, since there is still a high chance of winning through game updates only.

5 Discusion and Future Work

The above results show that an adversary trained against a gullible agent performs significantly worse against an agent with a more sophisticated technique of detecting deception based on logical contradictions between dialogue moves. This motivates the need for re-training the Learning Agent with these advanced adversaries using Reinforcement Learning (Rieser and Lemon, 2011). We will first target the case where the adversary can only detect lies within the same trading phase (Case 1), which we found to have the main impact on the agents' winning rates. We will present full results at the conference.

Acknowledgments

The research leading to this work has received funding from the European Community's ERC programme under grant agreement no. 269427 (STAC). The authors would like to thank Oliver Lemon for his comments on the draft.

References

Ioannis Efstathiou and Oliver Lemon. 2014. Learning non-cooperative dialogue behaviours. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 60–68, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari, editors. 2012. *EACL 2012: Proceedings of the Workshop on Computational Approaches to Deception Detection*, Avignon, France. Association for Computational Linguistics.

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

Verena Rieser and Oliver Lemon. 2011. *Reinforcement Learning for Adaptive Dialogue Systems*. Theory and Applications of Natural Language Processing. Springer.

Jaeeun Shim and Ronald C. Arkin. 2013. A taxonomy of robot deception and its benefits in HRI. In *SMC*, pages 2328–2335.

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

User Satisfaction *without* Task Completion

Peter Wallis
Centre for Policy Modelling
Manchester Metropolitan University
All Saints Campus, Oxford Road
Manchester, M15 6BH, UK
pwallis@acm.org

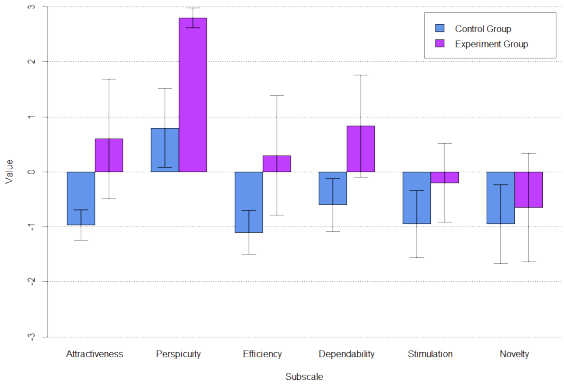


Figure 1: User satisfaction as measured using the UEQ questionnaire (Laugwitz et al., 2008) for two spoken dialog systems, one of which has added **social intelligence**.

Figure 1 presents an interesting result. We implemented a spoken dialog system that provided directory assistance, gave some subjects a set of tasks to do using the system, and then measured their user satisfaction with a standard HCI tool. We then gave the system some 'social intelligence' and re-ran the trial with the same tasks and with subjects from the same population. The graph shows a result that is not only significant (0.95) but also dramatic in that the experience was generally seen as positive, contrasting with the reaction to the control case and with popular opinion of IVR systems in general. A good result but, more importantly, note the experimental system did not work any better. The tasks were chosen such that only 20% were achievable no matter how good the interface. It seems better user satisfaction can be attained without making the system work better.

In previous papers we have given the background and motivation for our notion of social intelligence (Wallis, 2013), and described in detail the experimental setup (Wallis et al., 2014). In this short paper we describe the work underway to implement something more than a demonstrator.

1 The Theory

The idea that a computer could understand and use language has been with us from the very beginnings of computer science. Despite massive effort and considerable commercial potential, developments in the area have met with limited success. Historically the focus has been on the information conveyed by language but we are developing the idea that language is primarily social in purpose and function. Rather than focus on language and meaning, we focus on issues such as power and distance, roles and obligations, all within the context of normative relations and human/cultural expectations.

Moving down a level, we embrace Tomasello's claim that human communication is **intentional** and **cooperative** (Tomasello, 2008). This move is however the cumulation of 15 years looking at language as action in a social setting ranging from work on politeness (Wallis et al., 2001) and abuse (de Angeli et al., 2005) to conversational strategies (Wallis, 2008) and engagement (Wallis, 2010). In summary the key to language *as humans use it* is their surprisingly effective (but hard to notice) skills at recognising the intent of others. To use an example from Dennett, seeing two children tugging at a teddy bear, the human observer will be quite certain they both *want* it (Dennett, 1987). Even if people don't actually reason in terms of beliefs, desires and goals, the intentional stance we take is how we think others think when we communicate with them and is hence key to the recipient design of our utterances.

What is more we humans are (socially) compelled to cooperate in the process. If we have reached the point of being *engaged* (Wallis, 2010) in a conversation with someone, then we work hard to *account for* (Seedhouse, 2004) what he or

she says. Some actions – especially classic speech acts – are intended to be interpreted, other communicative acts are just 'radiated' (such as smiles) and others, such as unconsciously scratching your nose, are just acts. A chimpanzee according to Tomasello, is perfectly capable of recognising intent, but has no social compunction to interpret actions as those of communication.

2 The Mechanism

Given this is the true nature of the 'language instinct,' there are two challenges for those who want to engineer better conversational agents. How do we create a conversational agent that can recognise the intent of its interlocutor, and what intentions should the agent have and when? The current work in this area at CPM is using a Belief Desire and Intention (BDI) architecture to implement a dialogue manager. BDI has been used for this many times before (Ardissono and Boella, 1998; Wallis et al., 2001; Kopp et al., 2005; Wong et al., 2007) and such use is often, it seems, conflated with Good Old Fashioned AI models of conversation based on planning (Allen et al., 1995). In the rest of this article I will use the term BDI to mean a Rao and Georgeff (Rao and Georgeff, 1995) style BDI system which does not *do* planning but rather selects and manages plans from a static plan-library. Such architectures were introduced to explicitly address the issue of situated action associated with traditional planning systems; the advantage it has over more popular approaches to the issue such as Behaviour Based Robotics (Arkin, 1998) is that it maintains a notion of working to a recipe. A BDI architecture in the sense used here explicitly balances reactive and deliberative behaviour, managing plans rather than creating them.

Intention recognition is a task that poses some interesting and challenging problems for AI research but not all it poses are insurmountable. A large slab of the general problem can be handled using a BDI architecture and treating intention recognition as a variant of plan selection (Heinze, 2003). No doubt a human would do it better, but the interactive nature of the dialog problem means that, as long as the system can account for its failings in an understandable way, the human will forgive it in much the same way we accommodate children *without blame* for their lack of knowledge.

A bigger challenge is the question of what the system ought to intend (to do) and when. Ultimately of course machine learning should be able to mimic human (intentional) behaviour in a social setting and so identifying explicit intentions would become redundant. The problem is that unacknowledged theory tends to be embedded in the training data (Hovy, 2010). The recent Dialog State Tracking Challenge (Hen, 2014) is, although an excellent and exciting development, a case in point with notions of dialog state being based on Information State Update (Kreutel and Matheson, 2000). As a model of human communication ISU puts, we believe, too much emphasis on the information 'carried' (Reddy, 1993) by speech acts and pays insufficient attention to the larger structures within dialog we think of as intentional (Wallis, 2008).

Studying language has of course been the work of many for centuries if not millennia but such work tends to be seen as 'unscientific' by many with a physical sciences background and confined to the dusty shelves of forgotten libraries. Once one has a model of intention however, descriptions of people wanting X, believing Y, and Z being normative become concrete enough to implement. Creating models of causality in such relations is hard however because it is all so *obvious* to us humans - too obvious to notice. It takes special skills and training to do the noticing and what is needed is some cross disciplinary work to identify a set of intentional structures (defined as a BDI plans) that might be used by a synthetic social actor filling a particular set of roles. We have had past successes with researchers from Applied Linguistics using the ethnomethodological variant of Conversation Analysis (ten Have, 1999) and with Grounded Theory (Urquhart et al., 2010), but being at core computer scientists, we are open to suggestions.

3 Conclusion

The system used to produce the data in Figure 1 was a demonstrator that only worked for the tasks given to the subjects and used 'canned' expressions much like the classic chat-bot mechanism (Ali, 2001). Our aim now is to implement the system fully and deploy it with members of the public with real information needs.

References

2001. The ALICE artificial intelligence foundation, January. http://206.184.206.210/alice_page.htm.

J. F. Allen, L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. G. Martin, B. W. Miller, M. Poesio, and D. R. Traum. 1995. The TRAINS project: A case study in defining a conversational planning agent. *Journal of Experimental and Theoretical AI*, 7(7):7–48.

L Ardissono and G. Boella. 1998. An agent architecture for NL dialog modeling. In *Artificial Intelligence: Methodology, Systems and Applications*, Berlin. Springer (LNAI 2256).

Ronald C. Arkin, editor. 1998. *Behavior-Based Robotics*. MIT Press, Cambridge, MA.

Antonella de Angeli, Sheryl Brahnam, and Peter Wallis (eds). 2005. *Abuse: the darker side of Human-Computer Interaction*. INTERACT, Rome, September. http://www.agentabuse.org/.

Daniel C. Dennett. 1987. *The Intentional Stance*. The MIT Press, Cambridge, MA.

Clinton Heinze. 2003. *Modelling Intention Recognition for Intelligent Agent Systems*. Ph.D. thesis, Department of Computing and Information Systems.

2014. Dialog state tracking challenge 2&3. http://camdial.org/ mh521/dstc/.

Eduard Hovy. 2010. Injecting linguistics into nlp by annotation, July. Invited talk, ACL Workshop 6, NLP and Linguistics: Finding the Common Ground.

Stefan Kopp, Lars Gesellensetter, Nicole Kramer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide - design and evaluation of a real-world application. In *5th International working conference on Intelligent Virtual Characters*. http://iva05.unipi.gr/index.html.

Jörn Kreutel and Colin Matheson. 2000. Modelling dialogue using multiple inferences over information states. In *Proceedings of ICOS-2, 2nd Workshop on Inference in Computational Semantics*, Dagstuhl.

B. Laugwitz, T. Held, and M. Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In A. Holzinger, editor, *USAB 2008 (LNCS 5298)*, pages 63–76.

A. Rao and M. Georgeff. 1995. BDI agents: from theory to practice. Technical Report TR-56, Australian Artificial Intelligence Institute, Melbourne, Australia.

Michael J. Reddy. 1993. The conduit metaphor: A case of frame conflict in our language about language. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press.

Paul Seedhouse. 2004. *The Interactional Architecture of the Language Classroom: A Conversation Analysis Perspective*. Blackwell, September.

Paul ten Have. 1999. *Doing Conversation Analysis: A Practical Guide (Introducing Qualitative Methods)*. SAGE Publications.

Michael Tomasello. 2008. *Origins of Human Communication*. The MIT Press, Cambridge, Massachusetts.

C. Urquhart, H. Lehmann, and M. Myers. 2010. Putting the theory back into grounded theory: Guidelines for grounded theory studies in information systems. *Information Systems Journal*, 20(4):357–381.

Peter Wallis, Helen Mitchard, Damian O'Dea, and Jyotsna Das. 2001. Dialogue modelling for a conversational agent. In Markus Stumptner, Dan Corbett, and Mike Brooks, editors, *AI2001: Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence*, Adelaide, Australia. Springer (LNAI 2256).

Peter Wallis, Keeley Crockett, and Clare Little. 2014. When things go wrong. In Sarvapali D. Ramchurn, Joel Fisher, Avi Rosenfeld, Long Tran-Thanh, and Kobi Gal, editors, *Human-Agent Interaction Design and Models (HAIDM)*, Paris.

Peter Wallis. 2008. Revisiting the DARPA communicator data using Conversation Analysis. *Interaction Studies*, 9(3), October.

Peter Wallis. 2010. A robot in the kitchen. In *ACL Workshop WS12: Companionable Dialogue Systems*, Uppsala.

Peter Wallis. 2013. The intentional interface. In Mario Conci, Virginia Dignum, Mathias Funk, and Dirk Heylen, editors, *Proceedings of the Workshop on Computers as Social Actors*. CEUR, September. http://ceur-ws.org/Vol-1119/.

A. Wong, A. Nguyen, and W.R. Wobcke. 2007. Robustness of a spoken dialogue interface for a personal assistant. In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 123–127.