SeineDial: 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)

Sarah Brown-Schmidt, Jonathan Ginzburg, Staffan Larsson

Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue

Sarah Brown-Schmidt, Jonathan Ginzburg, Staffan Larsson (eds.)

Université Paris-Diderot (Paris 7), Paris Sorbonne-Cité, September 2012

Sponsors

MARGUERITE: Il m'a assuré que tu n'avais jamais été amoureux.
JACQUES: Oh! pour cela il a dit vrai.
MARGUERITE: Quoi! Jamais de ta vie?
JACQUES: De ma vie.
MARGUERITE: Comment! à ton âge, tu ne saurais pas ce que c'est qu'une femme?
JACQUES: Pardonnez-moi, dame Marguerite.
MARGUERITE: Et qu'est-ce que c'est qu'une femme?
JACQUES: Une femme?
MARGUERITE: Oui, une femme.
JACQUES: Attendez . . .

Denis Diderot *Jacques le fataliste et son maître*

We are happy to present SemDial 2012 (SeineDial), the 16th annual workshop on the Semantics and Pragmatics of Dialogue. This year's workshop is hosted at Université Paris-Diderot, named for the great *encyclopédiste*, himself a great writer of dialogues. SeineDial continues the tradition of presenting high-quality talks and posters on dialogue from a variety of perspectives such as formal semantics and pragmatics, artificial intelligence, computational linguistics, and psycholinguistics.

38 submissions were received for the main session, and each was reviewed by three experts. 16 talks were selected for oral presentation; the poster session hosts many of the remaining submissions, together with additional submissions that came in response to a call for late-breaking posters and demos.

We are lucky to have three world famous researchers as invited speakers—Eve Clark, Geert-Jan Kruijff, and François Recanati. Each of these represents a broad range of perspectives and disciplines. We are sure that their talks will stimulate much interest and at least some controversy. Together with the accepted talks and posters we look forward to a productive and interactive conference.

We are grateful to the reviewers, who invested a lot of time giving very useful feedback, both to the program chairs and to the authors, and to members of the local organizing committee, Anne Abeillé, Margot Colinet, and Gregoire Winterstein for their hard work in helping to bring the conference to fruition.

We are also very grateful to a number of organizations, who provided generous financial support to SeineDial:

- CLILLAC-ARP, Université Paris-Diderot
- Laboratoire de Linguistique Formelle, Université Paris-Diderot
- The Laboratoire d'excellence LabEx-EFL (Empirical Foundations of Linguistics), Paris Sorbonne-Cité.
- La région Île de France, through their competitive scheme *Manifestations scientifiques en Île-de-France hors DIM.*

Sarah Brown-Schmidt, Jonathan Ginzburg, Staffan Larsson
September, 2012

Programme Committee

Sarah Brown-Schmidt (University of Illinois, Urbana Champaign, Co-chair)

Staffan Larsson (Gothenburg University, Co-chair)

- Jennifer Arnold (University of North Carolina)
- Ron Artstein (Institute for Creative Technologies, LA)
- Ellen Gurman Bard (Edinburgh University)
- Luciana Benotti (Universidad Nacional de Crdoba)
- Claire Beyssade (Institut Jean Nicod, Paris)
- Nate Blaylock (IHMC)
- Johan Bos (Groningen University)
- Susan Brennan (SUNY, Stonybrook)
- Mark Core (Institute for Creative Technologies, LA)
- Mariapaola D'Imperio (LPL, Aix en Provence)
- David Devault (Institute for Creative Technologies, LA)
- Myroslava Dzikovska (Edinburgh University)
- Jens Edlund (KTH, Stockholm)
- Heather Ferguson (University of Kent)
- Raquel Fernández (University of Amsterdam)
- Victor Ferreira (UC San Diego)
- Claire Gardent (LORIA)
- Kallirroi Georgila (Institute for Creative Technologies, LA)
- Eleni Gregoromichelaki (King's College, London)
- Anna Hjalmarsson (KTH, Stockholm)
- Amy Isard (Edinburgh University)
- Elsi Kaiser (University of Southern California)
- Andrew Kehler (UC San Diego)
- Ruth Kempson (King's College, London)
- Ivana Kruijff-Korbayova (DFKI, Saarbrcken)
- Alex Lascarides (Edinburgh University)
- Oliver Lemon (Herriot Watt University)

- Danielle Matthews (Sheffield)
- Gregory Mills (Edinburgh University)
- Benjamin Spector (Institut Jean Nicod, ENS)
- Aliyah Morgenstern (Université Paris 3)
- Chris Potts (Stanford University)
- Laurent Prévot (LPL, Aix en Provence)
- Matthew Purver (Queen Mary, University of London)
- Antoine Raux (CMU)
- Hannes Rieser (Bielefeld University)
- Verena Rieser (Herriot Watt University)
- David Schlangen (Bielefeld University)
- Gabriel Skantze (KTH, Stockholm)
- Benjamin Spector (Institut Jean Nicod, Paris)
- Matthew Stone (Rutgers)
- David Traum (Institute for Creative Technologies, LA)
- Nigel Ward (UTEP)

Organizing Committee

- Jonathan Ginzburg (chair)
- Anne Abeillé
- Margot Colinet
- Gregoire Winterstein

CONTENTS

Referential Coordination through Mental Files

François Recanati
Institut Jean-Nicod
Ecole Normale Superieure
29 rue d'Ulm, 75005 Paris, France
recanati@ens.fr
http://www.institutnicod.org

On the standard model, linguistic communication makes it possible for the hearer to entertain the thoughts expressed by the speaker, and what makes that possible is the fact that the thoughts in question are encoded in the speakers words. However, there are challenges both to the idea that communication results in the sharing of thoughts, and to the idea that it works by encoding the thoughts. After briefly reviewing the contextualist challenge, which targets the latter idea, I will turn to another challenge to the standard model, raised by singular thought.

What characterizes singular thoughts, and especially indexical thoughts (the paradigm case), is the fact that the modes of presentation through which one thinks of objects are context-bound and perspectival. Such modes of presentation are best construed as mental files exploiting (and presupposing) certain contextual relations to the reference. This raises the communication problem, first raised by Frege: if indexical thoughts are context-bound and relation-based, how is it possible to communicate them to those who are not in the same context and do not stand in the right relations to the object? Arguably, one has to give up the claim that communication involves thought sharing, in such cases.

Following Frege, I will appeal to an important distinction between linguistic and psychological modes of presentation. Psychological modes of presentation are thought ingredients, while linguistic modes of presentation are encoded. Psychological modes of presentation are perspectival and context-bound: they are mental files whose role is to store information one can gain in virtue of standing in certain contextual relations to the reference of the file, so they are available only to subjects who are appropriately situated vis vis the object. It follows that thoughts involving such modes of presentation are not shareable with subjects who are not in the right type of context. But linguistic modes of presentation are fixed by the conventions of the language and they are shared by all the language users. They are public and serve to coordinate mental files in communication by constraining them to contain the piece of information they encode. In this way communication takes place even though the indexical thoughts entertained by the speaker are, in some sense, private and cannot be shared by the audience. Communication no longer involves the replication of thoughts only their coordination.

In the last part of the talk I will apply the coordination model of communication to the referential use of definite descriptions, and I will discuss a key objection based on the distinction between semantic reference and speakers reference.

Optimal Reasoning About Referential Expressions

Judith Degen
Dept. of Brain and Cognitive Sciences
University of Rochester
jdegen@bcs.rochester.edu

Michael Franke
ILLC
Universiteit van Amsterdam
m.franke@uva.nl

Abstract

The *iterated best response (*IBR*) model* is a game-theoretic approach to formal pragmatics that spells out pragmatic reasoning as back-and-forth reasoning about interlocutors' rational choices and beliefs (Franke, 2011; Jäger, 2011). We investigate the comprehension and production of *referential expressions* within this framework. Two studies manipulating the complexity of inferences involved in comprehension (Exp. 1) and production (Exp. 2) of referential expressions show an intriguing asymmetry: comprehension performance is better than production in corresponding complex inference tasks, but worse on simpler ones. This is not predicted by standard formulations of IBR, which makes categorical predictions about rational choices. We suggest that taking into account quantitative information about beliefs of reasoners results in a better fit to the data, thus calling for a revision of the game-theoretic model.

1 Introduction

Reference to objects is pivotal in communication and a central concern of linguistic pragmatics. If interlocutors were ideal reasoners, speakers would choose the most convenient referential expression that is sufficiently discriminating given the hearer's perspective, while hearers would choose the referent for which an observed referential expression is optimal given the speaker's perspective. But it would be folly to assume that humans are ideal reasoners, so the question is: how much do interlocutors take each other's perspective into account when producing and interpreting referential expressions?

A lot of work has been dedicated to this issue. For example, computational linguists have investigated efficient and natural rules for generating and comprehending referential expressions (see Dale and Reiter (1995) and Golland et al. (2010) for work directly related to ours). Many empirical studies have addressed the more specific questions of whether, when and/or how, hearers take speakers' *privileged information* into account (Keysar et al., 2000; Keysar et al., 2003; Hanna et al., 2003; Heller et al., 2008; Brown-Schmidt et al., 2008). Also, eye-tracking studies in the visual-world paradigm have been used to investigate how *quantity reasoning* influences the interpretation of referential expressions (Sedivy, 2003; Grodner and Sedivy, 2011; Huang and Snedeker, 2009; Grodner et al., 2010). In recent work closely related to ours, Stiller et al. (2011) and Frank and Goodman (2012) proposed a Bayesian model of producing and comprehending referential expressions in a game setting similar to the kind we consider here. We will more closely compare these related approaches in Section 6. Despite these various efforts, it is still a matter of debate whether or to what extent interlocutors routinely consider each other's perspective.

In order to contribute to this question, we follow a recent line of experimental approaches to formal epistemology and game theory (Hedden and Zhang, 2002; Crawford and Iriberri, 2007) to investigate how much *strategic* back-and-forth reasoning speakers and hearers employ in abstract language games. The tasks we investigate translate directly to the kind of *signaling games* that have variously been used to account for a number of pragmatic phenomena, most notably *conversational implicatures* (see, e.g., Parikh (2001), Benz and van Rooij (2007) or Jäger (2008)). A benchmark model of idealized step-by-step reasoning, called *iterated best response (*IBR*) model*, exists for these games (Franke, 2011; Jäger, 2011). IBR makes concrete predictions about the depth of strategic reasoning required to "solve" different kinds of referential language games, so that by varying the difficulty of our referential tasks, it is possible to both: (i) test the predictions of IBR models of pragmatic reasoning and (ii) determine the extent to which speakers and hearers reason strategically about the use of referential expressions.

Our data shows that participants perform better at reasoning tasks that IBR predicts to involve fewer inference steps. This holds for comprehension and production. However, our data also shows an interesting asymmetry: comprehension performance is better than production in corresponding complex inference tasks, but worse on simpler ones. This is not predicted by standard formulations of IBR which makes categorical predictions about rational choices. However, it is predicted by a more nuanced variation of IBR that pays attention to the quantitative information in the belief hierarchies postulated by the model.

Section 2 introduces signaling games as abstract models of referential language use. Section 3 outlines the relevant notions of IBR reasoning. Sections 4 & 5 describe our comprehension and production studies respectively. Section 6 discusses the results.

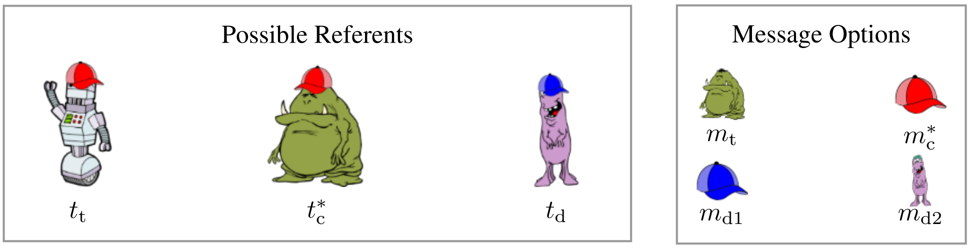
2 Referential Language Games

If speaker and hearer share a commonly observable set T of possible referents in their immediate environment, referential communication has essentially the structure of a *signaling game*: the sender S knows which $t \in T$ she wants to talk about, but the receiver R does not; the speaker chooses some description m; if R can identify the intended referent, communication is successful, otherwise a failure. Such a game consists of a set T (of possible referents), a set M of messages that S could use, a prior probability distribution $\Pr$ over T that captures R's prior expectation about the most likely intended referent, and a utility function that captures the players' preferences in the game. We assume that S and R are both interested in establishing reference, so that if t is the intended referent and t' is R's guess, then for some constants $s > f$: $\mathrm{U}(t, t') = s$ if $t = t'$ and f otherwise. Additionally, if messages are meaningful, this is expressed by a denotation function $[\![m]\!] \subseteq T$ that gives the set of referents to which m is applicable (e.g., of which it is true).

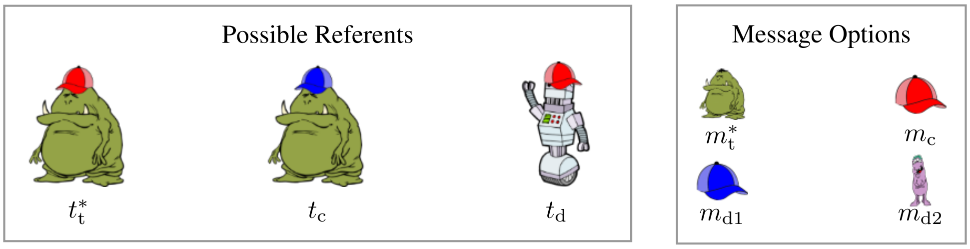
Consider, e.g., the situations depicted in Fig. 1. There are three possible referents $T = \{t_{\mathrm{t}}, t_{\mathrm{c}}, t_{\mathrm{d}}\}$ in the form of monsters and robots wearing one accessory each that both S and R observe. Since there is no reason to prefer any referent over another, we assume that $\Pr$ is a *flat* distribution over T. There are also four possible messages $M = \{m_{\mathrm{t}}, m_{\mathrm{c}}, m_{\mathrm{d1}}, m_{\mathrm{d2}}\}$ with some intuitively obvious "semantic meaning". For example, the message m_{c} for *red hat* would intuitively be applicable to either the *robot* t_{t} or the *green monster* t_{c}, so that $[\![m_{\mathrm{c}}]\!] = \{t_{\mathrm{t}}, t_{\mathrm{c}}\}$.

Signaling games like those in Fig. 1 are the basis for the critical conditions of our experiments (see also Sections 4 and 5), where we test which referent subjects choose for a given *trigger message* and which message they choose for a *trigger referent*. Trigger items for comprehension and production experiments are marked with an asterisk in Fig. 1. Indices t, c, d stand for *target*, *competitor* and *distractor* respectively.

We refer to a game as in Fig. 1(a) as the *simple implicature condition*, because it involves a simple *scalar implicature*. Hearing *trigger message* m_{c}^*, R should reason that S must have meant *target state* t_{t}, and not *competitor state* t_{c}, because if S had wanted to refer to the latter she could have used an unambiguous message. Conversely, when S wants to refer to *trigger state* t_{c}^*, she should not use the true but semantically ambiguous message m_{c}, because she has a stronger message m_{t}. Similarly, we refer to a game in Fig. 1(b) as the *complex implicature condition*, because it requires performing scalar reasoning twice in sequence (see Fig. 2 later on).

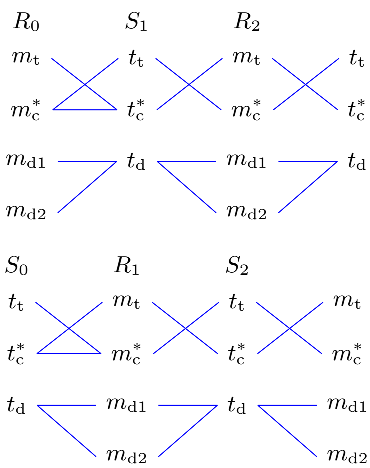


(a) simple

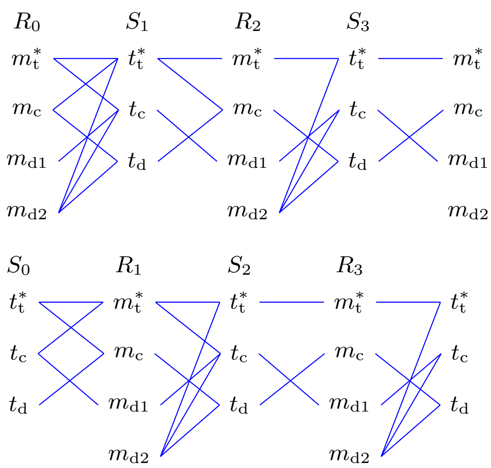


(b) complex

Figure 1: Target implicature conditions. Hearers choose one of the POSSIBLE REFERENTS $T = \{t_{\mathrm{t}}, t_{\mathrm{c}}, t_{\mathrm{d}}\}$. Speakers have MESSAGE OPTIONS $M = \{m_{\mathrm{t}}, m_{\mathrm{c}}, m_{\mathrm{d1}}, m_{\mathrm{d2}}\}$. Trigger items are indicated with asterisks: e.g., t_{t}^* is the referent to be communicated on complex production trials.



(a) simple



(b) complex

Figure 2: Qualitative predictions of the IBR model for simple and complex conditions. The graphs give the set of best responses at each level of strategic reasoning as a mapping from the left to the right.

3 IBR Reasoning

The IBR model defines two independent strands of strategic reasoning about language use: one that starts with a naïve (level-0) receiver R_0 and one that starts with a naïve sender S_0 (Franke, 2011; Jäger, 2011). If utilities are as indicated and priors are flat, the behavior of level-0 players is predicted to be a uniform choice over options that conform to the semantic meaning of messages: $R_0(m) = [\![m]\!]$ and $S_0(t) = \{m \mid t \in [\![m]\!]\}$. Sophisticated player types of level $k+1$ play any rational choice with equal probability given a belief that the opponent player is of level k. For our experimental examples, the "light" system of Franke (2011) applies, where sophisticated types are defined as:[1]

$$S_{k+1}(t) \quad = \begin{cases} \arg\min_{m \in R_k^{-1}(t)} \mid R_k(m) \mid \\ \qquad\qquad\qquad \text{if } R_k^{-1}(t) \neq \emptyset \\ S_0(t) \qquad\quad \text{otherwise} \end{cases}$$

$$R_{k+1}(m) = \begin{cases} \arg\min_{t \in S_k^{-1}(m)} \mid S_k(t) \mid \\ \qquad\qquad\qquad \text{if } S_k^{-1}(m) \neq \emptyset \\ R_0(m) \qquad\quad \text{otherwise} \end{cases}$$

The sequences of best responses for the simple and complex games from Fig. 1 are given in Fig. 2. On this purely qualitative picture, the IBR model makes the same predictions for comprehension and production. In the simple condition, the trigger item is mapped to either target or competitor with equal chance by naïve players; all higher level types map the trigger item to the target item with probability one. In the complex condition, the trigger items are mapped to target and competitor in levels 0 and 1 with equal probability, but uniquely to the target item for $k \geq 2$.

The sequences in Fig. 2 only consider the actual best responses of S and R, but not the more nuanced quantitative information that gives rise to these. Best responses are defined as those that maximize expected utility given what the players believe about how likely each choice option would lead to communicative success. The relevant expected success probabilities are given in Table 1 for sophisticated types. (Naïve types have no or only trivial beliefs about the game.)

For reasons of space suffice it to give the intuition behind these numbers. E.g., in the simple condition R_1 believes that the trigger message is used by naïve senders who want to refer to t_t or t_c. But naïve senders who want to refer to t_c would also use m_t with probability 1/2. So, by Bayesian conditionalization, after hearing m_c, R_1 believes the intended referent is t_t with probability 2/3.

Notice that while R's success expectations always sum to one (there is always only exactly one intended referent), S's success expectations need not (several messages could be believed to lead to successful communication). A further difference concerns when S and R are sure of communicative success. In the simple condition, S_1 is already sure of success, but only $R_{\geq 2}$ is. In the complex condition, R_2 is already sure of success, but only $S_{\geq 3}$ is. So, if we assume that human reasoners aim for certainty of communicative success in pragmatic reasoning, the simple condition is less demanding in production than in comprehension, while for the complex condition the reverse is the case.

4 Experiment 1

Exp. 1 tested participants' behavior in a *comprehension* task that used instantiations of the signaling games described in Section 2.

4.1 Methods

Participants. Using Amazon's Mechanical Turk, 30 workers were paid \$0.60 to participate. All were naïve as to the purpose of the experiment and participants' IP address was limited to US addresses only. Two participants did the experiment twice. Their second run was excluded.

Procedure and Materials. Participants engaged in a referential comprehension task. On each trial they saw three objects on a display. Each object differed systematically along two dimensions: its ontological kind (robot or one of two monster species) and accessory (scarf or either blue or red hat). In addition to these three objects, participants saw a pictorial message that they were told was sent to them by a previous participant whose job it was to get them to pick out one of these three objects. They

[1] Here $R_k^{-1}(t) = \{m \mid t \in R_k(m)\}$. Likewise for S_k^{-1}.

	simple		complex	
level	R	S	R	S
1	$\langle 2/3, 1/3, 0\rangle$	$\langle 1, 1/2, 0, 0\rangle$	$\langle 1/2, 1/2, 0\rangle$	$\langle 1/2, 1/2, 0, 1/3\rangle$
2	$\langle 1, 0, 0\rangle$	$\langle 1, 0, 0, 0\rangle$	$\langle 1, 0, 0\rangle$	$\langle 1/2, 0, 0, 1/3\rangle$
3	$\langle 1, 0, 0\rangle$	$\langle 1, 0, 0, 0\rangle$	$\langle 1, 0, 0\rangle$	$\langle 1, 0, 0, 1/3\rangle$

Table 1: Success expectations for the trigger items in the simple and complex condition. Success expectations for R are given in order for t_{t}, t_{c} and t_{d}, those for S in order for m_{t}, m_{c}, m_{d1} and m_{d2}.

were told that the previous participant was allowed to send a message expressing only one feature of a given object, and that the messages the participant could send were furthermore restricted to monsters and hats. The four expressible features were visible to participants at the bottom of the display on every trial.

Participants initially played four sender trials. They saw three objects, one of which was highlighted with a yellow rectangle, and were asked to click on one of four pictorial messages to send to another Mechanical Turk worker to get them to pick out the highlighted object. They were told that the other worker did not know which object was highlighted but knew which messages could be sent. The four sender trials contained three unambiguous and one ambiguous trial which functioned as fillers in the main experiment.

Participants saw 36 experimental trials, with a 2:1 ratio of fillers to critical trials. Of the 12 critical trials, 6 constituted a simple implicature situation and 6 a complex one as defined in Section 2 (see also Fig. 1).

Target position was counterbalanced (each critical trial occurred equally often in each of the 6 possible orders of target, competitor, and distractor), as were the target's features and the number of times each message was sent. Of the 24 filler trials, half used the displays from the implicature conditions but the target was either t_{c} or t_{d} (as identified unambiguously by the trigger message). This was also intended to prevent learning associations of display type with the target. On the other 12 filler trials, the target was either entirely unambiguous or entirely ambiguous given the message. That is, there was either only one object with the feature denoted by the trigger message, or there were two identical objects that were equally viable target candidates. Trial order was pseudo-randomized such that there were two lists (reverse order) of three blocks, where critical trials and fillers were distributed evenly over blocks. Each list began with three filler trials.

4.2 Results and Discussion

Proportions of choice types are displayed in Fig. 3(a). As expected, participants were close to ceiling in choosing the target on unambiguous filler trials but at chance on ambiguous ones. This confirms that participants understood the task. On critical implicature trials, participants' performance was intermediate between ambiguous and unambiguous filler trials. On simple implicature trials, participants chose the target 79% of the time and the competitor 21% of the time. On complex implicature trials, the target was chosen less often (54% of the time).

To test whether the observed differences in target choices above were significantly different, we fitted a logistic mixed-effects regression to the data. Trials on which the distractor was selected were excluded to allow for a binary outcome variable (target vs. no target choice). This led to an exclusion of 5% of the data. The model predicted the log odds of choosing a target over a competitor from a Helmert-coded CONDITION predictor, a predictor coding the TRIAL number to account for learning effects, and their interaction. Three Helmert contrasts over the four relevant critical and filler conditions were included in the model, comparing each condition with a relatively less skewed distribution against the more skewed distributions (in order: ambiguous fillers, complex implicatures, simple implicatures, unambiguous fillers). This allowed us to capture whether the differences in distributions for neighboring conditions suggested by Fig. 3(a) were significant. We included the maximal random effect structure that allowed the model to converge:[2] by-participant ran-

[2]For the procedure that was used to generate the random effect structure, see `http://hlplab.wordpress.com/`

	Coef β	SE(β)	z	p
(INTERCEPT)	1.81	0.22	8.3	<**.0001**
AMBIG.VS.REST	−2.56	0.45	−5.6	<**.0001**
COMPLEX.VS.EASIER	−3.20	0.53	−6.0	<**.0001**
SIMPLE.VS.UNAMBIG	−2.68	0.81	−3.3	<**.001**
TRIAL	0.00	0.01	0.3	0.8
TRIAL:AMBIG.VS.REST	−0.07	0.03	−2.6	<**.05**
TRIAL:COMPLEX.VS.EASIER	−0.01	0.03	−0.4	0.7
TRIAL:SIMPLE.VS.UNAMBIG	0.08	0.05	1.7	**0.08**

Table 2: Model output of Exp. 1. AMBIG.VS.REST, COMPLEX.VS.EASIER, and SIMPLE.VS.UNAMBIG are the Helmert-coded condition contrast predictors, in order.

dom slopes for CONDITION and TRIAL and by-item random intercepts. Results are given in Table 2.

All Helmert contrasts reached significance at $p < .001$. That is, all target/competitor distributions shown in Fig. 3(a) are different from each other. There was no main effect of TRIAL, indicating that no learning took place overall during the course of the experiment. However, there were significant interactions, suggesting selective learning in a subset of conditions. In particular there was a significant interaction between TRIAL and the Helmert contrast coding the difference between ambiguous fillers and the rest of the conditions (AMBIG.VS.REST, $\beta = -.05$, $SE = .02$, $p < .05$) and a marginally significant interaction between TRIAL and the Helmert contrast coding the difference between the simple implicature and unambiguous filler condition (SIMPLE.VS.UNAMBIG, $\beta = .08$, $SE = .05$, $p = .08$). Further probing the simple effects revealed that participants chose the target more frequently later in the experiment in the simple and complex condition. This was evidenced by a main effect of TRIAL on that subset of the data ($\beta = .03, SE = .01, p < .05$) but no interactions with condition. There were no learning effects in the ambiguous and unambiguous filler conditions; participants were at chance for ambiguous items and at ceiling for unambiguous items throughout. This suggests that at least some participants became aware that there was an optimal strategy and began to employ it as the experiment progressed.

We next address the question of whether the data supports the within-participant distributions predicted by standard IBR. Recall from Section 2 that for the simple condition, IBR predicts R_0 players to have a uniform distribution over target and competitor choices and $R_{\geq 1}$ players to choose only the target. For the complex condition, the uniform distribution is predicted for both R_0 and R_1 players, while only target choices are expected for $R_{\geq 2}$ players.

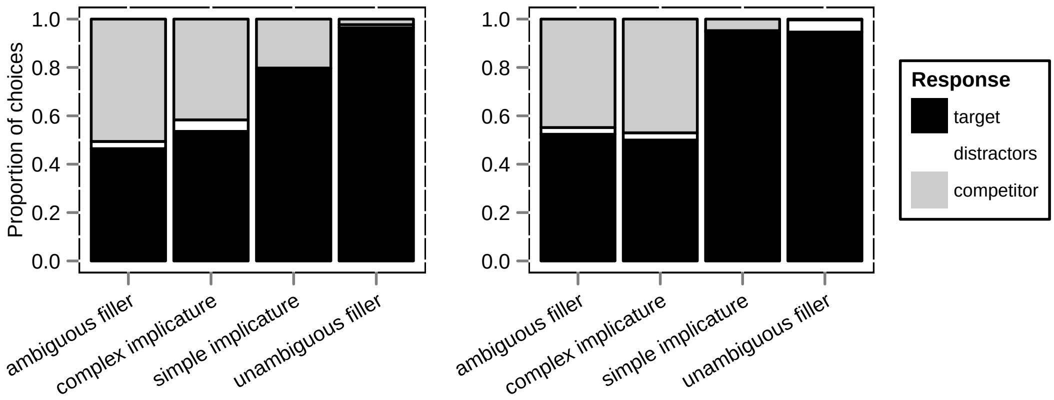
This is not borne out (see Fig. 4(a)). On the one hand, there were 3 participants in the simple condition and 5 in the complex condition who chose the target on half of the trials and could thus be classified as R_0 (or R_1 in the complex condition). Similarly, there were 11 participants in the simple condition and one in the complex condition who chose only targets and thus behaved as sophisticated receivers according to IBR. On the other hand, the majority of participants' distributions over target and competitor choices deviated from both the uniform and the target-only distribution.

One possibility is that some participants' type shifted from R_k to R_{k+1} as the experiment progressed. That is, they may have shifted from initially choosing targets and competitors at random to choosing only targets. However, while it is the case that overall more targets were chosen later in the experiment in both implicature conditions, there was nevertheless within-participant variation in choices late in the experiment inconsistent with a categorical shift. Another possibility is that the experiment was too short to observe this categorical shift.

5 Experiment 2

Exp. 2 tested participants' behavior in a *production* task that used instantiations of the signaling games described in Section 2.

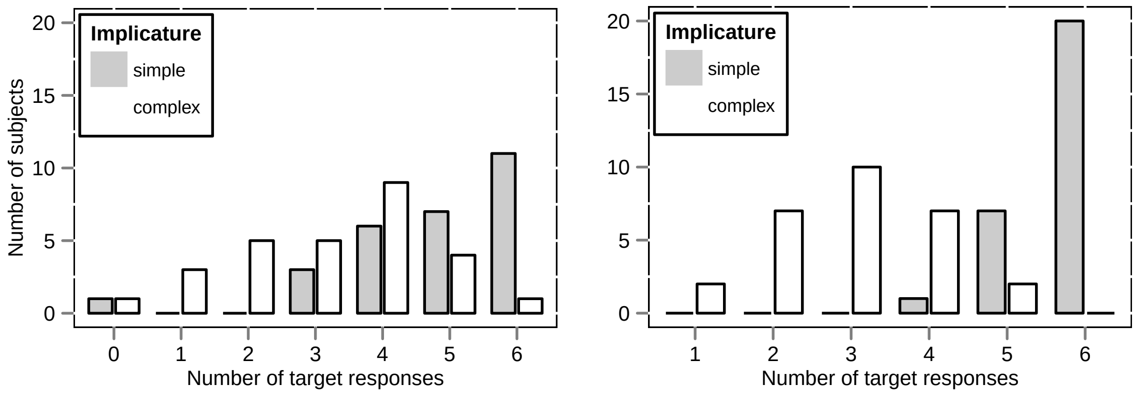
2009/05/14/random-effect-structure/



(a) Experiment 1

(b) Experiment 2

Figure 3: Proportions of target, competitor, and distractor choices in implicature and filler conditions (Exps. 1 & 2).



(a) Experiment 1

(b) Experiment 2

Figure 4: Distribution of participants over number of target choices in implicature conditions (Exp. 1 & 2).

5.1 Methods

Participants. Using Amazon's Mechanical Turk, 30 workers were paid $0.60 to participate under the same conditions as in Exp. 1. Data from two participants whose comments indicated that not all images displayed properly were excluded.

Procedure and Materials. The procedure was the same as on the sender trials in Exp. 1. Participants saw 36 trials with a 2:1 ratio of fillers to critical trials. There were 12 critical trials (6 simple and 6 complex implicature situations as in Fig. 1). Half of the fillers used the same displays as the implicature trials, but one of the other two objects was highlighted. This meant that the target message was either unambiguous (e.g. when the highlighted object was t_{t} in Fig. 1(a) the target message was m_{c}) or entirely ambiguous. The remaining 12 filler trials employed other displays with either entirely unambiguous or ambiguous target messages. Two experimental lists were created and counterbalancing was ensured as in Exp. 1.

5.2 Results and Discussion

Proportions of choice types are displayed in Fig. 3(b). As in Exp. 1, participants were close to ceiling for target message choices on unambiguous filler trials but at chance on ambiguous ones. On critical implicature trials, participants' performance was slightly different than in Exp. 1. Most notably, the distribution over target and competitor choices in the simple implicature condition was more skewed than in Exp. 1 (95% targets, 5% competitors), while it was more uniform than in Exp. 1 on complex implicature trials (50% targets, 47% competitors).

We again fitted a logistic mixed-effects regression model to the data. Trials on which the distractor messages were selected were excluded to allow for a binary outcome variable (target vs. competitor choice). This led to an exclusion of 2% of trials. In addition, the unambiguous filler condition is not included in the analysis reported here since there was only 1 non-target choice after exclusion of distractor choices, leading to unreliable model convergence. Thus, as in Exp. 1, CONDITION was entered into the model as a Helmert-coded variable but with only two contrasts, one comparing the simple implicature condition to the mean of ambiguous fillers and the complex implicature condition (SIMPLE.VS.HARDER), and another one comparing the ambiguous fillers with the complex implicatures (AMBIG.VS.COMPLEX). The model reported here further does not contain a TRIAL predictor to control for learning effects because model comparison revealed that it was not justified ($\chi^2(1) = 0.06$, $p = .8$). That is, there were no measurable learning effects in this experiment. We included the maximal random effects structure that allowed the model to converge: by-participant random slopes for CONDITION and by-item random intercepts.

The SIMPLE.VS.HARDER Helmert contrast reached significance ($\beta = 3.04$, $SE = 0.5$, $p < .0001$) while AMBIG.VS.COMPLEX did not ($\beta = 0.08$, $SE = 0.41$, $p = .9$). That is, there was no difference between choosing a target in the ambiguous filler condition and in the complex implicature condition, suggesting that participants were at chance in deriving complex implicatures in production. However, they were close to ceiling in choosing targets in the simple implicature condition.

The observed within-participant distributions are better predicted by the qualitative version of IBR than in Exp. 1 (see Fig. 4(b)). For the simple condition, IBR predicts S_0 players to have a uniform distribution over target and competitor choices and $S_{\geq 1}$ players to choose only the target. For the complex condition, the uniform distribution is predicted for both S_0 and S_1 players, while only target choices are expected for $S_{\geq 2}$ players.

In the simple implicature condition, 75% of participants were perfect S_1 reasoners. The remaining 25% chose almost only targets. That is, participants very consistently computed the implicature. In contrast, the bulk of participants chose targets versus competitors at random in the complex implicature condition. Only 2 participants chose the target 5 out of 6 times.

Comparing these results to the results from Exp. 1, we see the following pattern: in production the simple one-level implicatures are more readily computed than in comprehension, while the more complex two-level implicatures are more readily computed in comprehension than in production. That is, rather than comprehension mirroring production, in this paradigm there is an asymmetry between the two. This is consistent with the quantitative interpretation of IBR (as described in section 3) that takes into account players' uncertainty about communicative success.

6 General Discussion

In two studies using an abstract language game we investigated speakers' and hearers' strategic reasoning about referential descriptions. Most generally, our results clearly favor step-wise solution concepts like IBR over equilibrium-based solution concepts (e.g. Parikh (2001)) as predictors of participants' pragmatic reasoning: our results suggest that interlocutors do take perspective and simulate each others' beliefs, although (a) message and interpretation choice behavior is not always optimal and (b) perspective-taking decreases as the number of reasoning steps required to arrive at the optimal response, as predicted by IBR, increases.

We also found evidence for an intriguing asymmetry between production and comprehension. While not predicted by the standard formulation of the IBR model, this asymmetry is consistent with an interpretation of IBR that takes into account the uncertainty that interlocutors have about the probability of communicative success given a restricted set of message and interpretation options. This calls for a revision of the IBR model to incorporate more nuanced quantitative information. Since, moreover, there is a substantial amount of individual variation, further investigating the role of individual differences on perspective-taking (e.g. Brown-Schmidt (2009)) promises to be a fruitful avenue of further research that could inform model revisions.

It could be objected that the comparison of implicatures across experiments may be problematic due to the different nature of the tasks involved in the production vs. comprehension experiments and differences underlying the involved inference processes. However, note that the version of the IBR model that takes into account interlocutor uncertainty predicts the asymmetry between production and comprehension that we found precisely by integrating some of the differences involved in the two processes: most importantly, since conversation is modelled as a dynamic game, the sender reasons about the future behavior of the receiver, while the receiver reasons "backward", so to speak, using Bayesian conditionalization, about the most likely initial state the sender could have been in; this gives rise, as we have seen, to different predictions about when a speaker or a hearer can be absolutely certain of communicative success. How this difference is implemented mechanistically is an interesting question that merits further investigation.

Frank and Goodman (2012) report the results of an experiment using a referential game almost identical to ours and show that a particular Bayesian choice model very reliably predicts the observed data for both comprehension and production. In fact, the proposed Bayesian model is a variant of IBR reasoning that considers only a level-1 sender and a level-2 receiver, but assumes *smoothed* best response functions at each optimization step. In a smoothed IBR model, players' choices are stochastic with choice probabilities proportional to expected utilities (see Rogers et al. (2009) for a general formulation of such a model in game theoretic terms). This suggests a straightforward agenda for future work: combining our approach and that of Frank and Goodman (2012), smoothed IBR models that allow various strategic types for speakers and listeners should be further tested on empirical data.

In related work investigating comprehenders' capacity for deriving ad hoc scalar implicatures, Stiller et al. (2011) found that subjects could draw simple implicatures of the type we report above in a setup very similar to ours, but failed to draw complex ones. In contrast, our comprehenders performed above chance in the complex condition (albeit only slightly so). One possible explanation for this difference is that unlike Stiller et al. (2011), we restricted the set of message alternatives and also made it explicit to participants that a message could only denote one feature. This highlights the importance of (mutual knowledge of) the set of alternatives assumed by interlocutors in a particular communicative setting. While we restricted this set explicitly, in natural dialogue there is likely a variety of factors that determine what constitutes an alternative.

This suggests that future extensions of this work should move towards an artificial language paradigm. For example, whether a given message constitutes an alternative is likely to be affected by message complexity, which was held constant in our setup by using pictorial messages. Artificial language paradigms allow for investigating the effect of message complexity on inferences of the type reported here. Similarly, it will be important to further test the quantitative predictions made by IBR, e.g. by parametrically varying the payoff of communicative success and failure s and f and the interaction thereof with message complexity.

One question that arises in connection with the restrictions we imposed on the set of available pictorial messages, is the extent to which our results are transferable to natural language use. This is a legitimate concern that we would have to address empirically in future work. But notice also that, firstly, there is no *a priori* reason to believe that reasoning about natural language use and reasoning about our abstract referential games should necessarily differ — indeed it has been noted as early as Grice (1975) that conversational exchanges constitute but one case of rational communicative behavior. More importantly, even if reasoning about natural language *were* different in kind from strategic reasoning in general, the kind of strategic IBR reasoning we address here is a specific variety of reasoning that has been explicitly proposed in the literature as a model of pragmatic reasoning. The reported experiments are thus relevant in at least as far as they are the first empirical test of whether human reasoners are, in general, able to perform *this* kind of strategic reasoning in a task that translates the proposed pragmatic context models as directly as possible into an experimental setting.

We conclude that the studies reported are an encouraging first step towards validating game-theoretic approaches to formal pragmatics, which are well-suited to modeling pragmatic phenomena and generating quantitative, testable predictions about language use. The future challenge, as we see it, lies in fine-tuning the formal models alongside further careful empirical investigation.

Acknowledgements

We thank Gerhard Jäger, T. Florian Jaeger, and Michael K. Tanenhaus for fruitful discussion. This work was partially supported by a EURO-XPRAG grant awarded to the authors and NIH grant HD-27206 to Michael K. Tanenhaus.

References

Anton Benz and Robert van Rooij. 2007. Optimal assertions and what they implicate. *Topoi*, 26:63–78.

Sarah Brown-Schmidt, Christine Gunlogson, and Michael K. Tanenhaus. 2008. Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, 107:1122–1134.

Sarah Brown-Schmidt. 2009. The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin and Review*, 16(5):893 – 900.

Vincent P. Crawford and Nagore Iriberri. 2007. Fatal attraction: Salience, naïveté, and sophistication in experimental "hide-and-seek" games. *The American Economic Review*, 97(5):1731–1750.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Michael Franke. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics & Pragmatics*, 4(1):1–82.

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

H.P. Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Daniel Grodner and Julie C. Sedivy. 2011. The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter and E. Gibson, editors, *The Processing and Acquisition of Reference*. MIT Press, Cambridge, MA.

Daniel Grodner, Natalie M. Klein, Kathleen M. Carbary, and Michael K. Tanenhaus. 2010. "Some", and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116:42 – 55.

Joy Hanna, Michael K. Tanenhaus, and John C. Trueswell. 2003. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49:43–61.

Trey Hedden and Jun Zhang. 2002. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36.

Daphna Heller, Daniel Grodner, and Michael K. Tanenhaus. 2008. The role of perspective in identifying domains of reference. *Cognition*, 108:831–836.

Y. Huang and Jesse Snedeker. 2009. On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58:376–415.

Gerhard Jäger. 2008. Applications of game theory in linguistics. *Language and Linguistics Compass*, 2/3:406–421.

Gerhard Jäger. 2011. Game-theoretical pragmatics. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 467–491. Elsevier, Amsterdam.

Boaz Keysar, Dale J. Barr, and J. S. Brauner. 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11:32–37.

Boaz Keysar, S. Lin, and Dale J. Barr. 2003. Limits on theory of mind use in adults. *Cognition*, 89:25–41.

Prashant Parikh. 2001. *The Use of Language*. CSLI Publications, Stanford University.

Brian W. Rogers, Thomas R. Palfrey, and Colin Camerer. 2009. Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory*, 144(4):1440–1467.

Julie C. Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32:3–23.

Alex Stiller, Noah D. Goodman, and Michael C. Frank. 2011. Ad-hoc scalar implicature in adults and children. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Using a Bayesian Model of the Listener to Unveil the Dialogue Information State

Hendrik Buschmeier and **Stefan Kopp**
Sociable Agents Group – CITEC and Faculty of Technology, Bielefeld University
PO-Box 10 01 31, 33501 Bielefeld, Germany
{hbuschme, skopp}@uni-bielefeld.de

Abstract

Communicative listener feedback is a prevalent coordination mechanism in dialogue. Listeners use feedback to provide evidence of understanding to speakers, who, in turn, use it to reason about the listeners' mental state of listening, determine the groundedness of communicated information, and adapt their subsequent utterances to the listeners' needs. We describe a speaker-centric Bayesian model of listeners and their feedback behaviour, which can interpret the listener's feedback signal in its dialogue context and reason about the listener's mental state as well as the grounding status of objects in information state.

1 Introduction

In dialogue, the interlocutor not currently holding a turn, is usually not truly passive when listening to what the turn-holding interlocutor is saying. Quite the contrary, 'listeners' actively participate in the dialogue. They do so by providing communicative feedback, which, among other signals, is evidence of their perception, understanding and acceptance of and agreement to the speakers' utterances. 'Speakers' use this evidence to reason about common ground and to design their utterances to accommodate the listener's needs. This interplay makes communicative listener feedback an important mechanism for dialogue coordination and critical to dialogue success.

From a theoretical perspective, however, the interpretation of communicative feedback is a difficult problem. Feedback signals are only conventionalised to a certain degree (meaning and use might vary with the individual listener) and, as Allwood et al. (1992) argue, they are highly sensitive to their linguistic context – e.g., the speakers' utterances – and the communicative situation in general.

We present a Bayesian network model for interpreting a listener's feedback signals in their dialogue context. Taking a speaker-centric perspective, the model keeps representations of the mental 'state of listening' attributed to the listener in the form of belief states over random variables, as well as an estimation of groundedness of the information in the speaker's utterance. To reason about these representations, the model relates the listener's feedback signal to the speaker's utterance and his expectations of the listener's reaction to it.

2 Background and related work

Feedback signals, verbal-vocal or non-verbal, are communicative acts[1] that bear meaning and serve communicative functions. Allwood et al. (1992, p. 3) identified four *basic* communicative functions of feedback, namely *contact* (being "willing and able to continue the interaction"), *perception* (being "willing and able to perceive the message"), *understanding* (being "willing and able to understand the message"), and *attitudinal reactions* (being "willing and able to react and (adequately) respond to the message"). It is also argued that these functions form a hierarchy such that higher functions encompass lower ones (e.g., communicating understanding implies perception, which implies being in contact). Kopp et al. (2008) extended this set of basic functions by adding *acceptance/agreement* (previously considered an attitudinal reaction) and

[1]Note, however, that listeners might not be (fully) aware of some of the feedback they are producing. Not all should be considered as necessarily having communicative intent (Allwood et al., 1992). Nevertheless, even such 'indicated' feedback is communicative and is often interpreted by interlocutors.

by regarding expressions of emotion as attitudinal reactions

Feedback signals can likely take an infinite number of forms. Although verbal-vocal feedback signals, as one example, are taken from a rather small repertoire of lexical items such as ‘yes’, ‘no’, as well as non-lexical vocalisations such as ‘uh-huh’, ‘huh’, ‘oh’, ‘mm’, many variations can be produced spontaneously through generative processes such as by combination of different vocalisations or repeating syllables (Ward, 2006). In addition, these verbalisations can be subject to significant prosodic variation. Naturally, this continuous space of possible feedback signals can express much more than the basic functions described above. And listeners make use of these possibilities to express subtle differences in meaning (Ehlich, 1986) – which speakers are able to recognise, interpret (Stocksmeier et al., 2007; Pammi, 2011) and react to (Clark and Krych, 2004).

For a computational model of feedback production, Kopp et al. (2008) proposed a simple concept termed ‘listener state.’ It represents a listener’s current mental state of contact, perception, understanding, acceptance and agreement as simple numerical values. The fundamental idea of this model is that the communicative function of a feedback signal encodes the listener’s current mental state. An appropriate expression of this function can be retrieved by mapping the listener state onto the continuous space of feedback signals.

In previous work (Buschmeier and Kopp, 2011), we adopted the concept of listener state as a representation of a mental state that speakers in dialogue *attribute* to listeners through Theory of Mind. That is, we made it the result of a feedback interpretation process. We argued that such an ‘attributed listener state’ (ALS) is an important prerequisite to designing utterances to the immediate needs a listener communicates through feedback. The ALS captures such needs in an abstract form (e.g., is there a difficulty in perception or understanding) by describing them with a small number of variables, and is in this way similar to the “one-bit, most minimal partner model” which Galati and Brennan (2010, p. 47) propose as a representation suitable for guiding general audience design processes in dialogue.

For more specific adaptations, a speaker needs to consider more detailed information, such as the grounding status of previous utterances (Clark, 1996). Knowing whether previously conveyed information can be assumed to be part of the common ground (or even its degree of groundedness [Roque and Traum, 2008]) is important in order to estimate the success of a contribution (and initiate a repair if necessary) and to produce subsequent utterances that meet a listener’s informational needs.

Analysing an inherently vague phenomenon such as feedback signals in their dialogue context is almost only possible in a probabilistic framework. It is difficult to draw clear-cut conclusions from listener feedback and even human annotators, not being directly involved in the interaction, have difficulties consistently annotating feedback signals in terms of conversational functions (Geertzen et al., 2008).

A probabilistic framework well suited for reasoning about knowledge in an uncertain world is that offered by Bayesian networks. They represent knowledge in terms of ‘degrees of belief’, meaning that they do not hold one definite belief about the current state of the world, but represent different possible world states along with their probabilities of being true. Furthermore, Bayesian networks make it possible to model the relevant influences between random variables representing different aspects of the world in a compact model. This is why they are potentially well suited for reasoning about feedback use in dialogue. Using a Bayesian network, the conditioning influences between dialogue context, listener feedback, ALS, as well as the estimated grounding status of speaker’s utterances can be captured in a unified and well-defined probabilistic framework.

Representing grounding status not only in degrees of groundedness but also in terms of degrees of belief, adds a new dimension to the approach put forth by Roque and Traum (2008). Dealing with uncertainty in the representation of common ground simplifies the interface to vague information gained from listener feedback, and removes the need to prematurely commit to a specific grounding level. This keeps the information status of an utterance open to change.

Bayesian networks have already been used to model problems similar to the one in question. Paek and Horvitz (2000), for example, use Bayesian networks to manage the uncertainties, among other things, in the model of grounding behaviour in the ‘Quartet’ architecture for spoken dialogue systems. Rossignol et al. (2010) on the other hand created a Bayesian network model of dialogue system users' grounding behaviour. There the Bayesian network simulates consistent user behaviour which can be used for experimentation with, and training of, dialogue management policies. Finally, Stone and Lascarides (2010) propose to combine Bayesian networks with the logic based Segmented Discourse Representation Theory (SDRT; Asher and Lascarides, 2010) for a theory of grounding in dialogue that is both rational (in the utility theoretic sense) and coherent (by assigning discourse relations a prominent role in making sense of utterances).

3 A Bayesian model of the listener

A speaker's Bayesian model of a listener should relate dialogue context, listener feedback, the attributed listener state as well as the grounding status of the speaker's utterances to each other. Constructing such a model either needs corpora with fine-grained annotations of all these aspects of dialogue (to 'learn' it from data) or detailed knowledge about the relations (to design it). Apart from the fact that adequate corpora are practically non-existent, structure-learning of a Bayesian network can only infer conditional independence between variables and not their underlying causal relations. The top-ranking results of a structure learning algorithm might therefore differ substantially, resulting in networks that disagree about influences and causal relationships (Barber, 2012). For this reason, we take the approach of constructing a Bayesian network by 'hand', making – as is not uncommon in cognitive modelling – informed decisions based on research findings and intuition.

3.1 Assumed causal structure

When analysing or modelling a phenomenon with Bayesian networks, it is helpful to think of them as representing the phenomenon's underlying causal structure (Pearl, 2009). Network nodes represent causes, effects or both, and directed edges between nodes represent causality. A directed edge from a node *A* to a node *B*, for example, models that *A* is a cause for *B*, and that *B* is an effect of *A*. Another directed edge from *B* to a third node *C*, makes *B* the cause of *C*. Being intermediate, it is possible that *B* is both an effect (of *A*) and a cause (of *C*).

Figure 1 illustrates the causal structure of listener feedback in verbal interaction that we assume. In a given situation, a speaker *S* produces an utterance in the presence of a listener *L* and wants to know what *L*'s mental state of listening is towards her utterance, i.e., whether *L* is in contact, has perceived, understood and accepts or agrees with *S*'s utterance. As it is impossible for *S* to directly observe *L*'s mental state, she can only try to reconstruct it based on *L*'s communicative actions (i.e., *L*'s feedback) and by relating it to the dialogue context: her utterance, her expectations and the communicative situation.

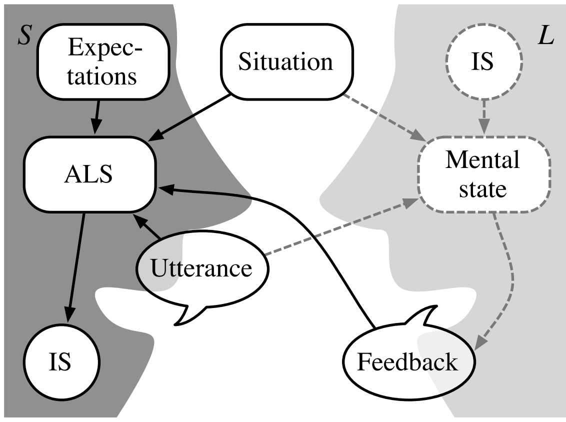


Figure 1: Speaker *S* reasoning about the mental state of listener L. *S*'s utterances cause *L* to move into a certain state of understanding. This influences *L*'s feedback signals, which are evidence for *S*'s attributed listener state of *L*.

To make a causally coherent argument, we assume, for the moment, that *L*'s unobservable mental state is part of the Bayesian listener model (parts unobservable to *S* are drawn with grey dashed lines in Figure 1). *L*'s mental state results from the effect of *S*'s utterance, the communicative situation as well as *L*'s information state. *L*'s mental state, on the other hand, causes him to provide evidence of his understanding by producing a feedback signal. In this way closure is achieved for the causal chain from utterance, via mental state and feedback signal, to *S*'s reconstruction 'ALS' of *L*'s mental state.

This causally coherent model can easily be reduced to an agent-centric model for *S*, which consists of only those influences that *S* can observe directly (drawn with black solid lines in Figure 1). Although this leads to a 'gap' in the causal chain, nodes retain their roles as causes and/or effects.

It should be noted, however, that the causal model only provides the scaffolding of a more detailed model to be presented next. Each node is a mere place-holder for a complete network structure. These sub-networks are constructed according to information that is available and useful to model feedback interpretation for a speaker.

3.2 Attributed Listener State

The core of the Bayesian model of the listener is the reconstruction of the listener's mental state, the attributed listener state. As described in Section 2, the model should give an estimate of whether the listener is in contact, how well she perceives and understands what the speaker says and to which degree the listener accepts and agrees to the utterance's content. As in previous models of (attributed) listener state (Kopp et al., 2008; Buschmeier and Kopp, 2011) the notions of contact, perception, understanding, acceptance and agreement are modelled with one variable each. Here, their values C, P, U, AC and AG, however, should be interpreted in terms of 'degrees of belief' instead of in terms of strength (which is modelled in terms of the variables' states – see Section 4.1).

The influences among the ALS variables are modelled after Allwood et al. (1992)'s hierarchy of feedback functions and Clark (1996)'s ladder of actions: perception subsumes contact, understanding subsumes perception and contact, acceptance and agreement subsume understanding perception and contact. This means, for instance, that if understanding is assumed, perception and contact can be assumed as well. A lack of perception, on the other hand, usually implies that understanding cannot be assumed. Thus, the influences are the following: C influences P, P influences U, and U influences AC and AG (see the central part of Figure 2 for a graphical depiction).

3.3 Contextual influences on ALS

The most important information for inferring the ALS is the listener's feedback signal itself. Thus, if it is recognised as having the communicative function 'understanding', there is a positive influence on the variables C, P and – especially – U. Variables AC and AG on the other hand are negatively influenced since speakers usually signal feedback of the highest function possible (Allwood et al., 1992; Clark, 1996).

To take into account the context-sensitivity of feedback signals, features of the speaker's utterance need to be considered in ALS estimation as well. If for example the speaker's utterance is simple[2], the degree of belief in the listener's successful understanding of the utterance should be high – even if explicit positive feedback is absent.

A further influence on ALS variables is how certain the listener seems to be about his mental state. A feedback signal can imply that a listener is still in the process of evaluating the speaker's statement – and is not yet sure whether she agrees with it – often by lengthening the signal or being hesitant of its production (Ward, 2006). This uncertainty could also influence the ALS.

Finally, situation specific influences and the influence of a speaker's expectations about the listener's behaviour are often connected to the dialogue domain and to known preferences in the listener. In a calendar assistant domain, which is the task domain we are working with, when presented, e.g., with a tight schedule and a new appointment of low priority, the likelihood is high that a listener rejects this new appointment.

3.4 Influences on Information State

The ALS mediates between the contextual factors described above and the information state. This makes the grounding status of the objects in the information state conditionally independent of the multitude of possible influencing factors which reduces the model's complexity significantly.

Each of the ALS variables influences the grounding status variable to a different degree. Believing that the listener is in full contact but neither perceives nor understands what the speaker is saying, for example, should lead to a low degree of belief in the groundedness of the object. In contrast, assuming the listener to have at least some understanding might be enough to consider information to be sufficiently grounded.

This part of the model can be considered one element of the speaker's 'grounding criterion' (Clark, 1996). The influences between ALS and information state map the listener's mental state (inferred from evidence of understanding) to groundedness of objects in information state. Whether the amount of groundedness is then considered 'sufficient for current purposes' (another element of the grounding process) is to be determined elsewhere.

[2]The notion of 'simplicity' is complex in itself. Here it is assumed that an utterance is simple if *(i)* it is not unexpected by the listener, *(ii)* it does not contains much new information and *(iii)* it is short.

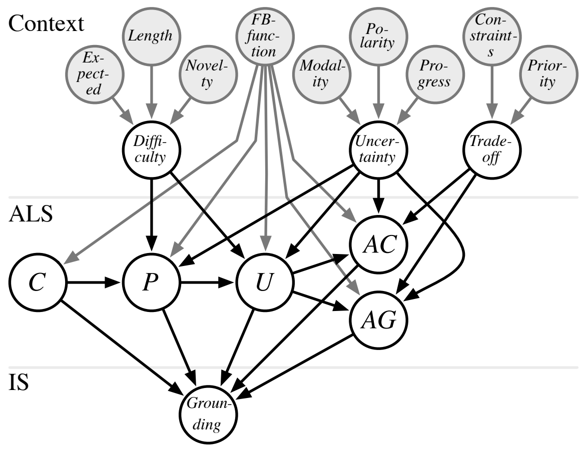


Figure 2: Structure of the Bayesian model of the listener. The variables shaded in grey are fully observable to a speaker (*FB function, modality, polarity,* and *progress* are derived from the listener's feedback signal).

4 Formal definition

We will now present the complete formal definition of the Bayesian model of the listener[3].It consists of a network structure, the node-internal structure, including their states, and parameters.

4.1 Model and node-internal structure

Figure 2 shows the structure of the full Bayesian network model of the listener. It reflects the causal structure sketched in Section 3.1 and Figure 1, and shows how the ALS sub-network, described in Section 3.2, acts as a layer mediating between context and information state.

Context itself consists of evidence nodes (drawn in shades of grey in Figure 2) that are directly observable to a speaker, and nodes for abstract concepts such as difficulty of the speaker's utterance, uncertainty of the listener, and the trade-off that the speaker expects the listener to make.

The node *Difficulty* reflects properties of the speaker's utterance and is part of the dialogue context. As described in footnote 2, it is an abstraction of utterance *Length* (having the states *short*, *medium*, *long*), of how *Expected* the utterance is (*low*, *medium*, *high*) as well as the *Novelty* of the information that is encoded in the utterance (*new*, *old*). *Difficulty* itself has the states *low*, *medium* and *high*. It influences the variables *P* and *U* in the ALS.

[3]A machine readable specification in the standardised 'Bayesian network interchange format' (XBIF) is available from `http://dx.doi.org/10.6084/m9.figshare.94357`.

The nodes *FB-function* and *Uncertainty* reflect properties of the listener's feedback signal. It is assumed that the communicative function of the listener's signal is classified externally and then represented in the node *FB-function*. This node can take the states *c, p, u, ac, ag,* $\neg c, \neg p, \neg u, \neg ac, \neg ag$, and *none*, which correspond to the basic functions as identified by Allwood et al. (1992) and Kopp et al. (2008). Feedback functions are distinguished according to their polarity (e.g., understood [u] versus not-understood [$\neg u$]). If the listener did not provide feedback, the state *none* might be chosen. The variable *FB-function* directly influences each of the ALS-variables.

Uncertainty is an abstract concept derived from the *Polarity* of the feedback signal (*positive*, *neutral* or *negative*), whether the signal conveys that the listener is still in *Progress* evaluating what the speaker uttered (*ongoing, finished*), and the *Modality* used to give feedback (*verbal, non-verbal, multimodal*). For example, a setting where *Polarity* is *neutral*, only one *Modality* is used, and *Progress* is *ongoing*, results in a degree of belief where the listener's uncertainty is *high*. The listener's uncertainty has an influence on the ALS-variables *P*, *U*, *AC* and *AG*.

Trade-off is an example of a domain-specific node that reflects the speaker's domain knowledge and his expectations of the listener's behaviour in the calendar assistant domain that we are using. It should not be considered to be an integral part of a general model of a listener. The trade-off a listener is expected to address depends on how many *Constraints*, i.e., other appointments a proposed appointment potentially interferes with (*none, one, a few, many*) and the *Priority* of the new appointment as compared to the priorities of the constraining appointments (*lower, similar, higher*). *Trade-off* itself can be *low*, *medium* and *high* and influences the variables *AC* and *AG* in the ALS.

Each of the ALS variables has the three states *low*, *medium*, and *high*. The variable *Grounding* with five states *low*, *low-medium*, *medium*, *medium-high* and *high* is more fine-grained and reflects a simple model of degrees of grounding (Roque and Traum, 2008). In general, both the ALS variables as well as the *Grounding* variable could be modelled with higher or lower number of states, and even as continuous random variables. Table 1 gives an overview of all variables/nodes and their states.

Table 1: Variables and their states in the Bayesian model of the listener. 'Meta nodes' correspond to the nodes described in Section 3.1 and displayed in Figure 1.

Meta nodes	**Variables**	**States**
ALS	*Contact*	*low, medium, high*
	Perception	*low, medium, high*
	Understanding	*low, medium, high*
	ACceptance	*low, medium, high*
	AGreement	*low, medium, high*
Utterance	*Difficulty*	*low, medium, high*
	–Expectable	*low, medium, high*
	–Length	*short, medium, long*
	–Novelty	*new, old*
Feedback	*–FB-function*	*none, c, p, u, ac, ag,* $\neg c, \neg p, \neg u, \neg ac, \neg ag$
	Uncertainty	*low, medium, high*
	–Modality	*verbal, non-verbal multimodal*
	–Progress	*ongoing, finished*
	–Polarity	*negative, neutral, positive*
Expectations	*Trade-off*	*low, medium, high*
	–Constraints	*none, one, a few, many*
	–Priority	*lower, similar, higher*
Inform. state	*Grounding*	*low, low-medium, medium, medium-high, high*

4.2 Model parameters

An important advantage of Bayesian networks over other probabilistic modelling approaches is that through the structure of the model (i.e., assuming conditional independences) a large reduction in the number of model-parameters is possible. The structure of our model allows a reduction of the full joint probability distribution with 1.870.672.320 parameters to a factored distribution consisting of only of 5.287 parameters.

As estimating this much smaller number of parameters by hand is still a tedious and error-prone task, we generated the model's parameters from a 'structured representation' of the conditional probability tables $\mathrm{cpt}(X_a)$ for each variable/node X_a and its influencing variables $X_i \in \mathrm{parents}(X_a) = \{X_i, \dots X_{i+n}\}$ in the following way:

1. Set the strength of influence that each variable X_i exerts on X_a by defining a weight $w_i \in [0,1]$ so that $\sum_{k=i}^{i+n} w_k = 1$.

2. For each variable X_i and its states $x_{i_j} \in \mathrm{states}(X_i) = \{x_{i_1}, \dots x_{i_z}\}$ assign a value $\iota_{x_{i_j}} \in [-1,1]$. x_{i_j} influences X_a negatively if $\iota_{x_{i_j}} < 0$, positively if $\iota_{x_{i_j}} > 0$, and does not have an influence if $\iota_{x_{i_j}} = 0$.

3. Now, for each possible combination of states $(x_{i_j}, \dots, x_{i+n_j}) \in \{\mathrm{states}(X_i) \times \dots \times \mathrm{states}(X_{i+n})\}$, calculate its weighted influence $\mu(x_{i_j}, \dots, x_{i+n_j}) = \sum_{k=i}^{i+n} w_k \cdot \iota_{x_{k_j}}$.

4. For each state $x_{a_j} \in \mathrm{states}(X_a) = \{x_{a_1}, \dots x_{a_z}\}$, assign a value $o_{x_{a_j}} \in [-1,1]$. Similarly to the definition given in step 2 above, $o_{x_{a_j}}$ determines the influence each combination c from step 3 has on a state x_{a_j}. A natural assignment for a variable with states *low, medium* and *high* would be $X_{a_{low}} = -1; X_{a_{medium}} = 0; X_{a_{high}} = 1$.

5. Now for each entry in the conditional probability table $\mathrm{cpt}(X_a)$ calculate a preliminary value $\tilde{p}(x_{a_j} | x_{i_j}, \dots, x_{i+n_j}) = \mathcal{N}(o_{x_{a_j}}, \mu(x_{i_j}, \dots, x_{i+n_j}))$, where $\mathcal{N}(o_{x_{a_j}}, \mu(x_{i_j}, \dots, x_{i+n_j}))$ is the value of the Gaussian probability density function at $o_{x_{a_j}}$ and with mean $\mu(x_{i_j}, \dots, x_{i+n_j})$.

6. Finally, normalise $\mathrm{cpt}(X_a)$ column-wise to convert the values $\tilde{p}(x_{a_j} | x_{i_j}, \dots, x_{i+n_j})$ into probabilities $p(x_{a_j} | x_{i_j}, \dots, x_{i+n_j})$.

In summary, this method generates the conditional probability table for a variable X_a by defining weighted means for each combination of states of its influencing variables. These are then used as means for Gaussian probability density functions, from each of which values at points $o_{x_{a_j}}$ associated with the states of the variable X_a are calculated. These are then converted to probabilities and put in the CPT.

With this method, instead of having to define the complete CPTs manually, i.e., a number of $x_{\mathrm{CPT}} = |\mathrm{states}(X_a)| \cdot \prod_{k=i}^{i+n} |\mathrm{states}(X_k)|$ parameters for each variable, only $x_{\mathrm{SR}} = |\mathrm{parents}(X_a)| + |\mathrm{states}(X_a)| + \sum_{k=i}^{i+n} |\mathrm{states}(X_k)|$ parameters are needed to define this structured representation of a conditional probability table. The loss of expressiveness caused by the structured representation was not limiting for defining the model – on the contrary, with its 254 parameters, it allowed for a straightforward expression of the relationships between variables.

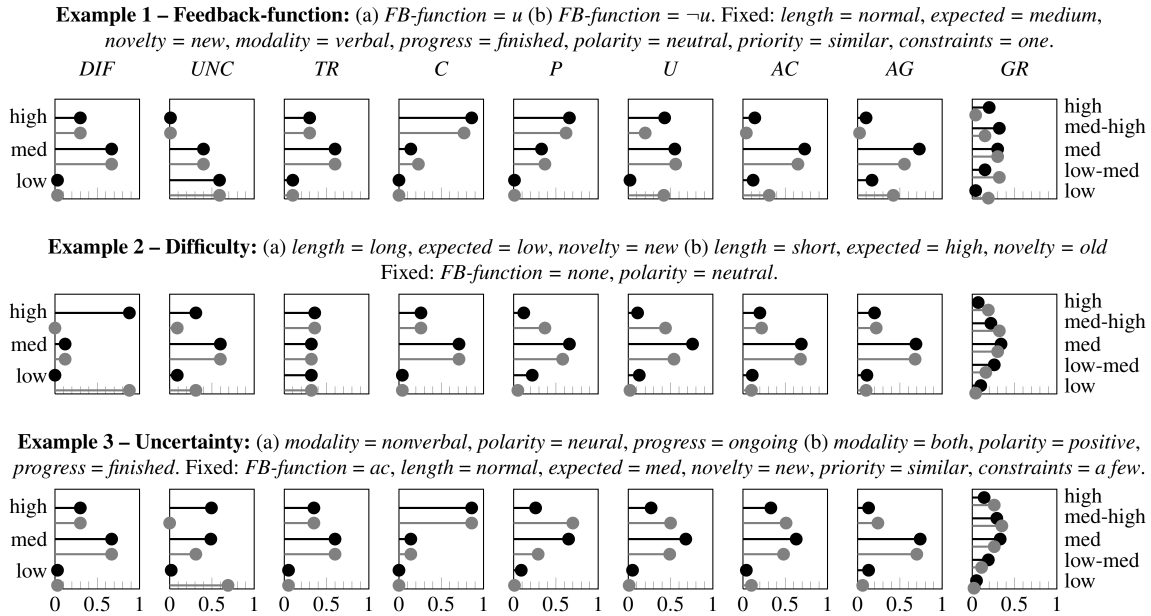


Figure 3: Plots of the belief states for three examples, each in two contrasting conditions. Conditions (a) are plotted with black, conditions (b) with grey comb lines. The x-axes show the degrees of belief of each variable's states. Variables are abbreviated as follows: *DIFficulty*, *UNCertainty*, *TRade-off*, *GRounding*. *C*, *P*, *U*, *AC* and *AG* are the ALS-variables.

5 Results

With the structure of the model defined, and the conditional probability tables generated from the structured representation, we use the Bayesian network and sensitivity analysis program SAMIAM[4] (Darwiche, 2009) to illustrate how the model behaves in some interesting situations. Figure 3 shows the belief states of the abstract context variables *Difficulty, Uncertainty* and *Trade-off*; the ALS-variables *C*, *P*, *U*, *AC* and *AG*; as well as the information state variable *Grounding*. The belief states are calculated given a certain fixed assignment of (some of) the variables representing the user's behaviour and the dialogue context. For each example, two contrasting belief states are displayed next to each other (conditions [a] drawn in black, conditions [b] in grey), reflecting the effect of a change in some variables while the others remain fixed.

Example 1, shows the influence a listener's feedback signal, in the form of its feedback function, has on ALS and grounding. It is assumed that the speaker will produce an utterance of normal length, that will not be unexpected, yet still contain new information. The belief state of the variable *Difficulty* (see Figure 3) indicates that this utterance will be of *medium* to *high* difficulty to the listener. It is further assumed that the listener either gives verbal feedback of function (a) understanding, e.g., 'uh-huh', or (b) non-understanding, e.g., 'huh' in response. The signal also conveys that the listener finished evaluating the utterance and thus, as the belief state of the variable *Uncertainty* indicates, seems to be rather certain about his evaluation. As a result, the belief states of all ALS variables show that feedback of type understanding in contrast to non-understanding results in a shift of the probability mass towards *medium* and *high* states. Similarly, for the variable *Grounding*, a higher degree of belief in groundedness of the utterance's content can be observed in the understanding condition (a).

Example 2 varies the difficulty of the speaker's utterance from (a) higher difficulty to (b) lower difficulty. The change in the evidence variables *Length, Expected* and *Novelty* is clearly reflected in the belief state of the variable *Difficulty*. It is assumed that the listener does not provide any feedback (i.e., *FB-function* is *none*). As a result, the probability mass in the belief states of the ALS variables *P* and *U* shift towards the *medium* and *low* states for the difficult utterance, and is more evenly distributed between the *medium* and *high* states for the simpler utterance. The same holds for the variable *Grounding*. The degree of belief

[4] `http://reasoning.cs.ucla.edu/samiam/`

in the utterance being grounded is higher for the simpler utterance. Notably, the belief states of the variables C, AC and AG are almost not affected. Utterance difficulty does not have a large impact on the listener being in contact, his acceptance of, or agreement with the utterance.

In Example 3 the listener responds to an utterance about an appointment which overlaps with a few other appointments (*Constraints* = *a few*) all of similar priority (*Priority* = *similar*). In both conditions, the listener communicates acceptance – but with different levels of uncertainty. In (a) the feedback signal is provided non-verbally, with neutral polarity and an indication that the listener's evaluation process is still ongoing (e.g., a hesitant and lengthened 'okay'). The belief state of the variable *Uncertainty* is mostly distributed between *medium* and *high*. In (b) feedback is provided both verbally and non-verbally, with a positive polarity and evidence that the evaluation is finished (e.g, a head nod in combination with an acknowledging 'okay'). Here the probability mass of *Uncertainty* is mostly distributed among the states *low* and *medium*. As a result, the belief states of the ALS variables for these two conditions differ for the variables P, U, AC and for AG (though only slightly). Although acceptance is communicated in both cases, higher uncertainty of the listener results in a shift of probability mass towards *medium* states instead of *medium* and *high* states. This also holds for the degree of belief in the utterance being grounded.

For each example the influences of variable changes on the belief states might seem small, but they might nevertheless make a significant difference in a decision theoretic process that operates on these probabilities. It should also be noted that the communicative situation was never impaired severely or even approached a breakdown. In general, the model parameters were chosen in such a way that negative feedback is required to make the *low* states of the ALS-variables likely, i.e., the model is optimistic about the listener's ability and willingness to perceive, understand, accept, and agree with what the speaker communicates.

6 Discussion and conclusion

Listener feedback is crucial for speaker–listener coordination in dialogue as it provides rich and subtle cues of the listener's mental state, as well as of the grounding status of information. We have presented a Bayesian network model for interpreting listener feedback for exactly these issues. It is important to note that the details of the model presented here should be regarded as just one concrete instantiation of a Bayesian model of listeners, and that we certainly did not (nor did we aim to) integrate everything that could influence the interpretation of feedback.

Nevertheless, our first modelling results reveal a number of interesting findings. Applying Bayesian networks enables a specification of the factors that contribute to the meaning of a feedback signal in a coherent, well-defined and interpretable formalism. Using this formalism, our model allows for direct reasoning about a listener's mental state, given certain evidence of perception, understanding, acceptance and agreement as provided by the listener in form of feedback, as well as the dialogue context. Built into the formalism is the capability to use the model diagnostically, i.e., reasoning from (assumed or asserted) listener states to possible feedback signals that most probably signal those. This can, for example, be used by the speaker to infer what kind of listener feedback would be most helpful under a particular uncertain dialogue situation. Having an idea of which kind of feedback is useful at the moment opens up the opportunity to produce a specific cue for the listener.

While reasoning about the listener's mental state and the groundedness of information, the model considers dialogue context in the form of a speaker's utterance and the speaker's expectations of the listener's reaction to the utterance. However, this must certainly be extended. For example, in a referential communication scenario, the situation could be modelled in terms of visibility and saliency of referents; in a noisy environment, the noise level could have an influence on the probability of an utterance being perceived and understood. Dialogue context could also be modelled in more sophisticated ways, for example by considering speech acts, and the ambiguity of the speaker's utterance.

An advantageous property of the model is its compatibility with incremental processing of feedback and incremental grounding in spoken dialogue systems. The model is constructed to run in parallel to a system's incremental output generation and, therefore, can influence the system behaviour even while it is being generated and synthesised (Buschmeier et al., 2012). Furthermore, the model is able to leverage subtle information about the listener's progress in processing the speaker's utterance, modulated, e.g., prosodically onto the feedback signal. It should be noted here, however, that the model currently does not regard temporal and discourse relationships – apart from the trivial relation that an utterance is followed by a feedback signal – in dialogue. Our plan is to make the model dynamic, taking influences of dialogue history and previous listener state on feedback interpretation into consideration (Stone and Lascarides, 2010).

Finally, using Bayesian networks makes it possible to adjust parameters to specific needs, even automatically and incrementally through learning. As described earlier, feedback signals are only conventionalised to a certain degree. It is likely that their usage and meaning differs between individual listeners. Currently, our model does not consider this, but idiosyncratic feedback meaning of listeners can easily be modelled via the model's structure and parameters. This bears the potential to make listener's idiosyncrasies 'transparent' and our Bayesian model of a listener can thus serve as a good starting point for studying the listener specific semantics and pragmatics of communicative feedback behaviour.

Acknowledgements This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence EXC 277 in '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.

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

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

Hendrik Buschmeier and Stefan Kopp. 2011. Towards conversational agents that attend to and adapt to communicative user feedback. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, pages 169–182, Reykjavik, Iceland.

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*, pages 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.

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

Adnan Darwiche. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, Cambridge, UK.

Konrad Ehlich. 1986. *Interjektionen*. Max Niemeyer Verlag, Tübingen, Germany.

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.

Jeoren Geertzen, Volha Petukhova, and Harry Bunt. 2008. Evaluating dialogue act tagging with naive and expert annotators. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1076–1082, Marrakech, Morocco.

Stefan Kopp, Jens Allwood, Karl Grammar, Elisabeth Ahlsén, and Thorsten Stocksmeier. 2008. Modeling embodied feedback with virtual humans. In Ipke Wachsmuth and Günther Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, pages 18–37. Springer-Verlag, Berlin, Germany.

Tim Paek and Eric Horvitz. 2000. Conversation as action under uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 455–464, Stanford, CA.

Sathish Pammi. 2011. *Synthesis of Listener Vocalizations. Towards Interactive Speech Synthesis*. Ph.D. thesis, Naturwissenschaftlich-Technische Fakultät I, Universität des Saarlandes, Saarbrücken, Germany.

Judea Pearl. 2009. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Antonio Roque and David R. Traum. 2008. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, Columbus, OH.

Stéphane Rossignol, Olivier Pietquin, and Michel Ianotto. 2010. Simulation of the grounding process in spoken dialog systems with Bayesian Networks. In *Proceedings of the 2nd International Workshop on Spoken Dialogue Systems Technology*, pages 110–121, Gotemba, Japan.

Thorsten Stocksmeier, Stefan Kopp, and Dafydd Gibbon. 2007. Synthesis of prosodic attitudinal variants in German backchannel "ja". In *Proceedings of Interspeech 2007*, pages 1290–1293, Antwerp, Belgium.

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

Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14:129–182.

The Pragmatics of Aesthetic Assessment in Conversation

Saul Albert
saul.albert@eecs.qmul.ac.uk
Patrick G.T. Healey
pat.healey@eecs.qmul.ac.uk

School of Electronic Engineering
and Computer Science
Queen Mary University of London

Abstract

Judgements of taste are intrinsic to everyday conversational interactions. People make assessments, agree and disagree, and negotiate these judgements as a core part of how they participate in activities, create and share knowledge, and manage their relationships with one another. This paper proposes a 'conversational aesthetics' that sees aesthetic assessments in terms of the pragmatics of talk-in-interaction.

A review of the literature of conversation analysis (CA) picks out conversational devices people routinely use when making assessments. These devices then inform the analysis of a transcribed conversation presented (unanalysed) in Anita Pomerantz' foundational 1984 paper on conversational assessment (Pomerantz, 1984) that deals with apparently aesthetic issues, in this case, judgements of taste about artworks.

This analysis suggests that people accomplish aesthetic judgements using the same generalized and ordinary mechanisms of conversational assessment that are ubiquitous in everyday talk. Analysing Pomerantz' data in terms of more recent developments in CA also poses some interesting methodological questions, and suggests further research into how people can offer up multiple parameters for judgement in aesthetic assessments, and how this process may involve shifts and step-wise drifts between conversational topics.

1 Introduction

Art historical approaches to aesthetics have conventionally treated judgements of taste as by-products of specific formal or perceptual qualities of the objects being judged (Greenberg, 1939), as circumscribed by historical and institutional conventions (Kristeller, 1951) (Danto, 1964) (Dickie, 2004), or as constituted by the societal and interpersonal relationships entailed between those involved through their participation in aestheticised spaces, objects or cultural contexts (Bourriaud, 2002), (Kester, 2004). However, as Michael Corris, a conceptual artist from the 1960s group *Art & Language* has pointed out, "such social effects are generally demonstrated rhetorically" (Corris, 2006), rather than with reference to any specific forms of evidence and analytical methods.

Recent ethnomethodological studies have analysed naturalistic video recordings of people in galleries and museums to demonstrate how their attention to, and thereby experiences of artworks are constituted through their interpersonal interactions (Lehn, 2006). However, these studies have focused almost exclusively on people's movements, gestures and physical orientation; partly because of the practical challenges of recording conversations in galleries (Hindmarsh et al., 2002), and partly to remedy a perceived imbalance in favour of using interviews and surveys in the field of visitors studies (Lehn and Heath, 2001).

Building on this approach to aesthetics as an essentially interactional activity, but looking at interactions outside the specific institutional context of the art gallery or museum, this paper draws on the methods of conversation analysis (CA) to ask what everyday judgements of taste look like in terms of the analysis of talk-in-interaction. The literature of CA is reviewed here in order to identify key conversational devices people routinely use in everyday assessments. These devices are then used as a toolkit to analyse a naturalistic conversation about an artwork.

2 Mechanisms of conversational assessment

An obvious example of the ubiquity of judgements of taste in conversation are the everyday assessments of the weather that Erving Goffman terms 'small talk' or 'safe supplies' of chat, readily available to neutralise the potentially offensive situation of 'painful silence' (Goffman and Best, 1982).

In order to develop a CA-informed approach to aesthetic assessments, the following introduction to CA mechanisms and methods highlights low-level conversational devices people use when making these kinds of routine judgements of taste.

2.1 Sequence and turn-taking

Goffman's sociological approach sees these 'supportive' or 'remedial' exchanges as interactional norms with which speakers state and reinforce their social relationships (Goffman and Manning, 2009). Sacks (1987) builds on this basis to identify the apparatus used by participants to coordinate their conversational exchanges as sequences of utterances organised in elegantly interlocking 'turns' (Sacks and Schegloff, 1974a) bound together by frequently observable 'conventional parts' or 'adjacency pairs'(Schegloff and Sacks, 1969) such as exchanges of greetings. Schegloff (1968) shows how conventional parts of these pairs are evidently relevant to conversational participants by highlighting the way they work to overcome or mark problems in their talk that regularly appear when a relevant part of a pair is omitted.

For example, by showing how people deal with, account for and 'repair' (Schegloff et al., 1977), (Schegloff, 1992) misunderstandings, mishearings or 'attributable silences' (Stephen C. Levinson, 1983) at different turn positions in telephone conversations, Schegloff (2004) demonstrates various uses of the word "hello": to initiate a simple exchange of greetings, or as a second part response to the first part of a telephone ringing, or in the midst of a conversation as a presence indicator to resume an interrupted call.

In the same way, an apparent request for a personal assessment of wellbeing: "How are you", can be used as an initial greeting, as a reciprocation of an earlier mutual enquiry (Schegloff, 1968), (Heritage, 1998), as an occasion to draw the conversation to a close (Schegloff and Sacks, 1969), or as kind of conversation restart marker or topic shifter: a way to "move out of talk *about* a trouble" (Jefferson, 1985).

Focusing on terms that are demonstrably relevant to participants themselves (through regular production or marked omission), CA develops a mico-analysis of sequentially relevant utterances by cataloguing the methodical organisation of talk, and extrapolating types from that data (Sacks, 1987).

This approach self-consciously denies the relevance of any quantitative measures or theories of communication that are 'outside of talk' (Schegloff, 1993), relying instead on observing the work people do to make their talk and behaviour contingently and situatedly intelligible in the sense of 'observable-and-reportable' (Garfinkel, 1994). In this way, the relative preference for some methodical organisations of talk, and the dispreference for others emerges from systematic observations.

This notion of conversational preference is the basic building-block of CA required for developing a picture of how a CA-informed aesthetics might work.

2.2 Preference organisation and aesthetic assessments

Levinson (1983) distinguishes between the common definition of preference as an explicit wish from the technical sense of preference as the conversational path of least resistance: the one least marked by requests for clarification or repair from subsequent speakers.

This structure of preference and dispreference is one of the key analytical tools in CA because so much of what is relevant and available to conversational participants is only observable in the delays, pauses, softenings and deferrals that characterise the production of a dispreferred response, whereas agreement and contiguity is the hard-to-detect, unmarked norm (Sacks, 1987).

Combining analytical sensitivity to regular sequences and marked omissions, as well as dispreference and 'reluctance markers' (Bilmes, 1988) that characterise deferred or marked responses, CA uses preference organisation as a "formal and anonymous apparatus for agreement/disagreement" (Sacks, 1987) that is teleologically independent of conversational context and topic.

In the following example, Pomerantz (1984) demonstrates the preference organisation of an assessment in which an initial assessment is 'shaped' for agreement, i.e. designed in such a way as to orient towards an agreement-preferred response:

```
     A: God izn it dreary.
 →              (0.6)

     A:   [ Y'know I don't think-
 (D) B:     .hh- it's warm though,
```

1: An agreement-preferred initial assessment shape featuring assessment of the weather. (Pomerantz, 1984) (NB:IV:11.-1).

In (1), an agreement-preferred initial assessment shape indicates that whereas the preferred response would be for immediate agreement (Sacks, 1987), the dispreferred disagreement indicated here with a (D), is marked as problematic by an attributable silence (Stephen C. Levinson, 1983) of 0.6 seconds and a partial softening of the disagreement with the weakened agreement modifier: "though" (Pomerantz, 1984).

This example demonstrates how even small-talk is coordinated with the same preference apparatus of delays, pauses and softenings of dispreferred responses common to participation in and agreement and disagreement with any assessment, even once the topic of the conversation moves beyond initial setting talk (Sacks and Schegloff, 1974b), (Maynard, 1984). As later examples will demonstrate, the same low-level mechanisms used here to talk about the weather are used in similar ways in extended conversations about art.

To summarise, the CA notion of preference sees assessments as organised sequences of utterances produced in interlocking turns, used to defer, delay and/or soften the impact of dispreferred second parts, or to reinforce the contiguity and agreement of preferred seconds with respect to their prior turn shapes.

2.3 Socioepistemic authority

> "[W]ith an assessment, a speaker claims knowledge of that which he or she is assessing." (Pomerantz, 1984).

Building on the idea of the authoritative primacy of recounting first-hand experience (Sacks, 1984), Pomerantz shows how participation and epistemic authority to assess are indexed in the participants' own terms by the way speakers account for not assessing something. Where an initial assessment invites a second response, a second speaker will regularly account for their not producing a second assessment by claiming insufficient access to or knowledge of the thing being assessed (Pomerantz, 1984).

Heritage (2005) uses these analytical building blocks systematically to index differences in 'epistemic authority' in talk as people introduce and negotiate different topics for assessment in conversation. An initial assessment sets out an 'information territory' with associated epistemic rights for different participants, which may be modified, challenged, downgraded or confirmed by subsequent assessments.

These modifications can be pre-emptive, for example, a speaker can downgrade the epistemic authority of an initial assessment with the use of an evidential such as 'seems' to shape an assessment for subsequent modification (Heritage and Raymond, 2005). Similarly, the authority of assessments can be pre-emptively bolstered by shaping them in the strongest terms for an agreed response, for example using the negative interrogative tag-question: 'isn't it?' (Heritage, 2002) to constrain subsequent assessments.

Specific prefixes such as 'Oh' often indicate a state-change in information territory which Heritage (1998) observes people using systematically in a way that re-orients the temporal or topical state of the conversation towards a new information territory in which participants may claim more or less authority to assess. This is observable in conversations in which co-participants compete for epistemic priority, claiming or ceding information territory by systematically differentiating their positions on assessments, even when seemingly reaching agreement.

For example, in (2), *D* and *C* are being asked by *A* to offer an assessment of a newly acquired print: "D'yuh li:ke it?", after which a second assessment

```
   A: D'yuh li:ke it?
(+) D:  ˙hhh Yes I do like it=
(-) D: =although I rrealy::=
   C: =Dju make it?
   A: No We bought it, It's a  ˙hh a Mary Kerrida print.
   D: O:h (I k-)=
   A: =Dz that make any sense to you?
   C: Mn mh.  I don' even know who she is.
   A: She's that's, the Sister Kerrida,  who,
   D:                                      [˙hhh
   D: Oh[that's the one you to:ld[me you bou:ght.=
   C:   [Oh-                      [
   A:                             [Ye:h
```

(2) The evaluation of a new artwork from (JS:I. -1) (Pomerantz, 1984, p.78).

becomes relevant. *C's* subsequent question, and disclaiming of any knowledge of the author ("I don' even know who she is.") soften and defer the dispreferred critical assessments (Pomerantz, 1984) indicated by the bracketed minus mark.

Applying Heritage's (1998) observations about how this process demarcates information territories to Pomerantz's example, both *C* and *D* use 'Oh' in this extract while differentiating their responses to *A's* question. Firstly, when *C* asks who made the print, and *A* explains who the author is, *D* replies using "O:h (I k–)=" possibly beginning to mark a different information territory from *C's*. Later, when *A* initiates a comprehension check: "=Dz that make any sense to you?", *C* explicitly disclaims knowledge of the author of the print ("I don' even know who she is."), accounting for the lack of a second assessment, demonstrating diminished epistemic rights to assess the print. Once again, *D* follows *A's* explanation about the author with another "Oh", marking a state-change and a subtle temporal shift from *A's* explanation about the author to a prior conversation between *D* and *A* about the print: "Oh that's the one you to:old me you bou:ght.=". Finally, *C* then uses an 'Oh', seemingly to acknowledge the differentiation.

This illustrates what Heritage (2005) characterises as "a systematic dilemma at the heart of agreement sequences" in which co-participants generally seek mutual agreement, but when providing it, "must respect the other party's information territories and associated epistemic rights".

It also demonstrates how the process of shifting between these territories by means of subtle temporal shifts and marked state-changes can be linked to shifts between conversational topics.

2.4 Topical shift

Conversational sequences are conventionally tied together into contiguous topics by questions or assessments being followed by responses on the same topic. However, topics evidently do change (Sacks, 1987), often by means of disjunctive topic-shift markers such as 'anyway', 'so', or 'Oh!' (Jefferson, 1984), (Maynard, 1984).

Sacks (1987) also observes unmarked 'stepwise' topic shifts, which Heritage and Atkinson (1984) describe as the aspect of conversation most "complex... and recalcitrant to systematic analysis". This may be partly due to the complexity of the coordination of minute overlaps in speech and rapid uses of acknowledgement tokens such as 'OK', 'Yeah', or 'mhmmm' that characterise step-wise topic shifts (Jefferson, 1981), making it hard to identify segues from one topic to the next: a kind of 'topical drift'. It may also be an inherent limitation of CA's methodological commitments: if the shift between topics is unmarked by participants, it may be unavailable for analysis.

If topics in conversation can be seen as information territories with different associations of epistemic rights for each participant, assessment sequences provide participants with specific conversational devices for moving between those territories, such as parameter shifts.

2.4.1 Parameter shift

```
A1  A: God izn it dreary.
                .
                .
A2  P:  ˙hh- it's warm though
```

(3) Pomerantz' (1984) example showing the contrastive assessment of "a shifted parameter", (NB:IV:11.-1).

In (3), Pomerantz revisits her earlier example of weather-talk to point out what she calls 'a shifted parameter' (Pomerantz, 1984) by which the weather is assessed, in this case marked by a "though". Here, the parameter of assessment shifts from the appearance of the weather, to the temperature. Pomerantz later expands on these kinds of shifts in parameter, and how they can start to modify the way participants refer to the things they are assessing.

Pomerantz uses example (4) in a footnote to highlight an unusual type of assessment that agrees with, then upgrades its prior, and then accomplishes what she calls a "subtle referent shift" (Pomerantz, 1984) of the upgraded assessment.

```
A: They look nice together.
B: Yes they're lovely.  But I
   particularly like the blue
   en gray, en [white,
A:             [Yeah
B: What's so nice about this is
   you get two nice pieces.
```

(4) A further example of "parameter-shift" in (JS:II:137) from Pomerantz (1984, p.98).

Here *A* assesses two vases, citing the parameter of their looking nice together. *B* initially agrees, even upgrading this assessment: "Yes they're lovely", then, marking the parameter-shift with a "But", slightly modifies the assessment to point out the colours. *A* concurs with an interjected acknowledgement token "Yeah". *B* then modifies the overall parameter of *A's* assessment: the niceness of the objects "together", instead assessing their colour and appearance, and finally emphasising that: "you get *two [distinct]* nice pieces".

In terms of Sacks' and Jefferson's distinction between disjunctive and step-wise shifts, these *parameter shifts* are marked by disjunctive 'but' or 'though' tags, deployed within an organisation of assessment sequences that tend towards overall agreement and contiguity, softening dispreferred disagreements by means of a subtle stepwise shift. In Heritage's terms, this could be seen as co-participants asserting the independence of their assessments by shifting between subtly differentiated information territories.

Analysing assessment sequences that display this type of topical drift via parameter shift is problematic as much of the CA apparatus for dealing with assessments and their epistemic territories depends on reliably reading second assessments as "produced by recipients of prior assessments in which the referents in the seconds are the same as those in the priors" (Pomerantz, 1984). An assessment that has undergone sufficient parameter shift to amount to a kind of topical drift could be seen and treated by conversational participants as a "fully sentential declarative assessment" (Heritage and Raymond, 2005). Alternatively, it may be seen as a grey area characterised by Jefferson (1981) in her analyses of the complexity of topic-shifting in phone conversations as an ambiguously attributed information territory in which the interactional cohesiveness of an exchange may be expressed as normal[1] while at the same time the conversation is undergoing an unmarked 'topical rupture' (Atkinson and Heritage, 1984).

The following section uses the CA methods and conversational devices outlined above to analyse a conversation about an artwork presented, but not analysed in any detail, in Pomerantz' paper on conversational assessment. (Pomerantz, 1984).

3 Conversational Aesthetic Assessment

Analysing Pomerantz's example of an conversation about judging an artwork demonstrates how the same conversational devices ubiquitous in everyday talk are present in extended aesthetic assessment sequences.

Informed by more recent developments in CA, this analysis also suggests how participants negotiate the assessment by shifting between various information territories within an overall conversational topic.

[1]Jefferson observes how this ambiguity about topical territory seems to engender exaggerated forms of recipient assessment feedback and affiliation such as "collaborative completion" (Jefferson, 1981).

```
     A: D'yuh li:ke it?
 (+) D:  ˙hhh Yes I do like it=
 (-) D: =although I rreally::=
     C: =Dju make it?
     A: No We bought it, It's a  ˙hh a Mary Kerrida print.
     D: O:h (I k-)=
     A: =Dz that make any sense to you?
     C: Mn mh.  I don' even know who she is.
     A: She's that's, the Sister Kerrida,  who,
     D:                                      [˙hhh
     D: Oh [that's the one you to:ld [me you bou:ght.=
     C:     Oh-                      [
     A:                              [Ye:h
     D:  [Ya:h.
     A:   Right.
        (1.0)
     A: It's worth something,
        (1.0)
     A: There's only a hundred of'm
        (0.5)
     D: Hmm
     E: Which picture is that.
     A: The one thet says Life.
        (1.5)
     A: (            ).
 (-) D: ˙hhh Well I don't- I'm not a great fan of this type of a:rt.  There are
        certain- ones I see thet I like, But I like the w- =
     E: =Is there ano[thuh way of spelling Life?.
 (-) D:              -more realistic-.
     A: hhmh!
     E: That's all [I wd loo(hh)k fo(h),
     D:             hh!
 (-) D: Yih d-know why I don't go fer this type of uh:  art, Becuz it- it
        strikes me ez being the magazine adverti:sement ty:pe.  Which some
        uh-uh some a' them are really great.  But tuhm I-my, taste in art is
        for the more uhit-t-treh- it tends tuh be realistic.
```

(5) Evaluation of a new artwork from (JS:I. -1) (Pomerantz, 1984). In Pomerantz' transcription, a '-' sign indicates critical assessment and a '+' sign marks an appreciative assessment.

3.1 Sequence, Turn and Preference Organisation

In (5), the same low-level organisation of sequences, turns and preference make the conversation amenable to a CA-informed analysis.

A first offers up a print for assessment: "D'yuh li:ke it?", after which a second assessment becomes relevant to all those addressed. *D* responds immediately with a token preferred affirmation upgraded by an emphatic "Yes I do like it=". However, *D's* final assessment: "I don't go fer this type of uh: art =although…", produced only after a long series of turns by multiple participants can be seen as presaged by the modifier token "although".

Possibly reacting to this marker of an immanent critical assessment, *C* interrupts[2] *D's*, offering up an alternative candidate parameter for assessing the print: its authorship ("=Dju make it?").

This interruption, as well as *E's* later interruption: "=Is there ano-thuh way of spelling Life?"[3] both take place just before *D* produces a critical

[2] Pomerantz (1984) uses Jefferson's CA transcription style in which an equals sign at the beginning or end of an utterance indicates an interruption or lack of a pause or gap between speaker turns, and square brackets stretching over one or more lines indicate overlaps (Atkinson and Heritage, 1984).

[3] It is unclear from the transcript whether *E's* question is a topic-relevant interjection or a side conversation.

assessment, suggesting a degree of spontaneous group coordination in the softening of *D's* dispreferred assessment, which remains relevant but deferred until the last seven turns.

Four long pauses of 0.5 - 1.5 seconds after *A* says "Right", mark the sustained absence of a second position assessment relevant to *A's* initial question, and indicate that these pauses can be read as attributable (even painful) silences (Goffman and Best, 1982).

D's ultimately critical assessments start out shaped in a manner directly counter posed to *A's* initial question: ("D'yuh li:ke it?") with "'hhh Well I don't-", which *D* softens somewhat with a weakened critical assessment: "I'm not a great fan", and a generalisation of the referent from *A's* specific print to "this type of a:rt.".

3.2 Indices of Epistemic Authority

This conversation demonstrates an intense negotiation over information territories (Heritage and Raymond, 2005), in which participants seem to compete over who has primary epistemic authority to assess the print on their own terms.

D's partial repeat and modification of the initial question: "D'yuh li:ke it?", with stress on the copula "do": "'hhh Yes I do like it=", is an example of what Tanya Stivers (2005) describes as a method of asserting primary rights from second position.

C's epistemic authority in the assessment of the print is first undermined by *A's* explicit comprehension check: "Dz that make any sense to you?", shifting referent from the object of the overall epistemic struggle (the print), to *C's* comprehension of *A's* prior turn, which *C* answers with an agreement tag "Mm mh", and then fails to produce a second assessment, accounting for this omission by claiming lack of knowledge or access: "I don't even know who she is".

D interrupts *A's* explanation of who the author is with several 'oh'-prefixed responses: "Oh that's the one you to:ld me you bou:ght.=", shifting to talk about a different time and a different conversation, possibly constituting a shift to an information territory that is differentiated from *C's* declination to assess the print.

This marked shift from the question of the authorship of the print to the subject of a prior conversation between *A* and *D*, functions as another deferral and also as a claim of *D's* epistemic authority to assess. These shifts between information territories are accompanied by shifts between parameters for an assessment.

3.3 Parameter Shift

In this conversation, participants offer up different parameters for assessment, withholding or shifting away from clearly critical, dispreferred second assessments.

For example, *C* interrupts *D* when offering up an important criterion for assessing the print: its authorship, and particularly, whether *A* themselves is the author. A series of turns follow in which the question of the authorship of the print functions as a backdrop to a rapid offering-up of multiple possible assessment criteria including:

- authorship,
- knowledge of the author,
- monetary value,
- scarcity,
- knowledge about the print,
- correct spelling,[4]
- how 'realistic' it is, and
- how much like a magazine advert it is.

After the initial discussion of authorship is concluded, a quick exchange of "Ye:h, Ya:h, Right" acknowledgement tokens between *A* and *D* marks readiness for a topic shift (Jefferson, 1984), which in this case is organised as a parameter-shift from the local assessment of the criterion of authorship, back to the deferred overall assessment of the print.

A then offers two further parameter-shifts, proposing new criteria for assessment in each subsequent turn: "It's worth, something," or "There's only a hundred of'm". Each of these short turns are marked with attributable silences of up to 1.5 seconds, that Maynard (1980) characterises as failed speaker transitions, marking attempts at topic shifts where further topical talk from others becomes relevant but in this case, remains unsatisfied when *A* themselves takes up the next turn again. *D* finally interjects with an emphatic "Hmm", marking *D's* turn to propose assessment criteria, starting with how realistic the print is.

D's dispreferred critical assessment is somewhat softened by *E's* concurrent interruption, pos-

[4]Assuming that *E's* interjection about the correct spelling of the word "Life" in the print relates to the print in question and is not a side conversation, see note[2].

sibly checking on a spelling within the print, and suggesting that the spelling is "all I ws loo(hh)k fo(h),", i.e. a relevant criterion for *E's* judgement of the print.

D's quick prespeech inbreath "˙hhh" (Jefferson, 1985) is followed by the delivery of a critical second assessment including the introduction of a further parameter for assessment, which functions as an account for the dispreferred response: the likeness of the print to a type of advertising.

4 Discussion

The CA-informed analysis of aesthetic assessments presented so far proposes conversational aesthetics as a distinctive set of interactional practices identifiable in the regular use of specific conversational devices. Future work may involve identifying more such devices as well as building on this initial analysis of how topical shift and step-wise drift are managed in aesthetic assessments.

4.1 Aesthetics as Information Territory

The conversation about *A's* print demonstrates assessments of taste in which the mechanisms of sequence, turn and preference organisation operate in each local assessment of various parameters, as well as in the overall global assessment via the deferral and softening of *D's* dispreferred second assessments.

Even after deferral, it seems that *D's* second assessments are softened to oblique critical assessments of "this type of a:rt", rather than a direct answer to *A's* initial question "D'yuh li:ke it?". *D's* conclusion: "I-my, taste in art is for the more uh:: uh it-t-treh- it tends tuh be realistic" is further softened by accounting for the assessment via a claim of "my taste", which functions here as the most explicit demarcation of an information territory to which *D* can claim exclusive access and absolute rights to assess. Even in this territory the assessment is still marked by agreement-disagreement assessment sequences ("some a' them are really great. But"), and tentative evidentials such as 'tends to be' (Heritage and Raymond, 2005).

In building up a picture of these core features of conversational aesthetic assessment, it is useful to bear in mind the CA view that any aesthetic discussion, however large or small, in any context, can be seen as a series of assessment sequences in which dispreferred second assessments are deferred by pauses, softenings and disjunctive parameter shifts.

4.2 Future Work: Parameter Drift

This paper has suggested how marked, disjunctive parameter shifts in which multiple candidate parameters are offered up for assessment can organise and facilitate a kind of step-wise topic shift. There are also unmarked drifts into different parameters and sub-topics such as *E's* apparent confusion about which print is being discussed segueing, unmarked, into either a separate side conversation about another print, or possibly into a discussion about whether the correct spelling of 'Life' in a print is a relevant parameter for its assessment.

The initial work presented here suggests that the nesting of local assessments with different parameters within an overall assessment may be seen as a landscape through which co-participants negotiate epistemic authority over their respective information territories. Further research is proposed into how, through this process, subtle topic shifts may be introduced into conversations, making new topics relevant to participants in the same way that in (5), the author of the print "Mary Kerrida" becomes relevant to the assessment.

This begs the question of whether shifts of topic via shifts in assessment parameters might help to explain how participants accomplish movement from topic to topic. The resulting availability of new and possibly unexpectedly relevant themes, contexts and objects for discussion might be useful as a pragmatic description of what could be called a creative conversation.

However, even if this idea is borne out by available conversational data, picking out shifts in the meanings and content that are brought into these interactions through assessment sequences rather than concentrating only on the structure and regularities of the shifts themselves stretches CA's methodological commitment to analysing only those interactions available to researchers and evidently at issue to the conversational participants.

Bearing the limitations of a CA-informed conversational aesthetics in mind, the next steps of this research will involve selecting naturalistic conversational data from the British National Corpus (BNC) (Burnard, 2000), re-transcribing some conversations in CA style using the newly published Audio BNC (Coleman et al., 2012) and analysing everyday dialogue sampled outside specific aesthetic/gallery contexts that demonstrates some of the conversational devices of aesthetic assessment discussed in this paper.

4.3 Conclusion

It may be inconsistent to move from a CA-informed structural analysis to an interpretative analysis by taking into account the ostensible meaning or the assessed content of conversational topics. However, the availability of taste-talk and art-talk in the CA literature itself is the basis for this initial attempt to describe the machinery of everyday conversational aesthetic assessments.

Treating aesthetic assessments as conversationally negotiated information territories highlights the difference between this approach and art-historical aesthetic theories that have tended to focus on the centrality of the work of art itself, either as mimesis, expression, form, narrative or conceptual content, or on its positioning within specific socio-political contexts or relations (den Braembussche, 2009).

An analysis of how these theories relate to this conversational aesthetic approach is beyond the scope of this paper, but it is worth pointing out that post-modernist theories of art and culture (Lyotard, 1984) and some related sociological aesthetic theories (Wolff, 1993) make compellingly similar observations about the social construction of art, as well as the ways in which art can be used to as a healthy and respectful outlet for discussion and dissonance within a society (Mouffe, 2002).

These theories are, however, methodologically at right-angles to a CA-informed approach to aesthetics. Although CA accounts of sequence, turn-taking, preference organisation, negotiated epistemic authority and step-wise topic shift may seem irrelevant to conventional aesthetic discourses, the material evidence available to CA enables a systematic interactional analysis of how people make aesthetic assessments in everyday speech, without prioritising specific formal, narrative or contextual norms other than those evidently relevant to conversational participants.

5 Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was funded by the RCUK Digital Economy Programme.

References

J.M. Atkinson and J. Heritage. 1984. *Structures of social action: Studies in conversation analysis*. Cambridge University Press, Cambridge.

Jack Bilmes. 1988. The concept of preference in conversation analysis. *Language in Society*, 17(2):161–181.

N Bourriaud. 2002. *Relational aesthetics*. Les presses du reel, Paris.

L Burnard. 2000. Reference Guide for the British National Corpus (World Edition).

John Coleman, Ladan Baghai-Ravary, John Pybus, and Sergio Grau. 2012. Audio BNC: the audio edition of the Spoken British National Corpus.

Michael Corris. 2006. The Dialogical Imagination : The Conversational Aesthetic of Conceptual Art. In David Hopkins, editor, *Avant Garde Critical Studies*, pages 301–310. Rodopi, Amsterdam.

Arthur Danto. 1964. The artworld. *The Journal of Philosophy*, 61(19):571–584.

Antoon Van den Braembussche. 2009. *Thinking Art*.

G Dickie. 2004. The new institutional theory of art. In Lamarque e Olsen, editor, *Aesthetics and the philosophy of art: the analytic tradition: an anthology.*, pages 47–54. Blackwell, Oxford.

Harold Garfinkel. 1994. *Studies in ethnomethodology*. Polity Press, Cambridge.

E Goffman and J Best. 1982. *Interaction Ritual: Essays in Face-to-face Behavior*. Pantheon, New York.

E Goffman and P Manning. 2009. *Relations in public: microstudies of the public order*. Transaction Publishers, New York.

Clement Greenberg. 1939. Avant-garde and kitsch. *Partisan Review*, pages 1–13.

J. Heritage and G. Raymond. 2005. The Terms of Agreement: Indexing Epistemic Authority and Subordination in Talk-in-Interaction. *Social Psychology Quarterly*, 68(1):15–38, March.

J. Heritage. 1998. Oh-prefaced responses to inquiry. *Language in Society*, pages 291–334.

John Heritage. 2002. The limits of questioning : negative interrogatives and hostile question content. *Journal of Pragmatics*, 34:1427–1446.

Jon Hindmarsh, Christian Heath, Dirk vom Lehn, and Jason Cleverly. 2002. Creating assemblies:. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW '02*, page 156, New York, New York, USA. ACM Press.

Gail Jefferson. 1981. On the articulation of topic in conversation. Technical report, Social Science Research Council, London.

G. Jefferson. 1984. Notes on a systematic Deployment of the Acknowledgement tokens 'Yeah' and 'Mmhm'. *Papers in Linguistics*, 17(2):197–216.

Gail Jefferson. 1985. On stepwise transition from talk about a trouble to inappropriately next-positioned matters. chapter 9, pages 191–222. Cambridge University Press, Cambridge.

GH Kester. 2004. *Conversation pieces: Community and communication in modern art*. University of California Press, Berkeley.

Paul Oskar Kristeller. 1951. The modern system of the arts: A study in the history of aesthetics part I. *Journal of the History of Ideas*, 12(4):496–527.

D Vom Lehn and Christian Heath. 2001. Exhibiting interaction: Conduct and collaboration in museums and galleries. *Symbolic interaction*, 24(2):189–216, May.

D Vom Lehn. 2006. Die Kunst der Kunstbetrachtung : Aspekte einer pragmatischen Ästhetik in Kunstausstellungen. *Soziale Welt*, 57:83–99.

J F Lyotard. 1984. *The Postmodern Condition: A Report on Knowledge*. Theory and History of Literature. Manchester University Press, Manchester.

DW Maynard. 1980. Placement of topic changes in conversation. *Semiotica La Haye*, 4(1980):263–290.

DW Maynard. 1984. Topical talk, ritual and the social organization of relationships. *Social Psychology Quarterly*, 47(4):301–316.

C. Mouffe. 2002. For an agonistic public sphere. In Okwui Enwezor, editor, *Democracy unrealized: documenta 11, platform 1*, pages 87–97. Hatje Cantz.

A Pomerantz. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J Maxwell Atkinson and John Heritage, editors, *Structures of social action: Studies in Conversation Analysis*, chapter 4, pages 57–102. Cambridge University Press, Cambridge.

H Sacks and EA Schegloff. 1974a. A simplest systematics for the organization of turn-taking for conversation. *Language*.

H Sacks and EA Schegloff. 1974b. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Harvey Sacks. 1984. On doing "being ordinary". In John Heritage and J Maxwell Atkinson, editors, *Structures of social action: Studies in Conversation Analysis*, chapter 6, pages 413–429. Cambridge University Press, Cambridge.

Harvey Sacks. 1987. On the preferences for agreement and contiguity in sequences in conversation. In G Button and J.R. Lee, editors, *Talk and social organization*, pages 54–69. Multilingual Matters, Clevedon.

EA Schegloff and Harvey Sacks. 1969. Opening up closings. *Contract*, 49(638).

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.

Emanuel Schegloff. 1968. Sequencing in Conversational Openings. *American Anthropologist*, (70):1075–1095.

Emanuel A Schegloff. 1992. Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *American Journal of Sociology*, 97(5):1295.

Emanuel A Schegloff. 1993. Reflections on Quantification in the Study of Conversation. *Research on Language & Social Interaction*, 26(1):99–128.

Emanuel Schegloff. 2004. Answering the Phone. In G H Lerner, editor, *Conversation Analysis: Studies from the First Generation*, chapter 2, pages 63–109. John Benjamins Publishing Company, Amsterdam.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge.

Tanya Stivers. 2005. Modified Repeats: One Method for Asserting Primary Rights From Second Position. *Research on Language & Social Interaction*, 38(2):131–158, April.

J Wolff. 1993. *The Social Production of Art*. New York University Press, New York, London.

A Cognitive Model for Conversation

Nicholas Asher
CNRS, Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, Toulouse
asher@irit.fr

Alex Lascarides
School of Informatics,
University of Edinburgh
alex@inf.ed.ac.uk

Abstract

This paper describes a symbolic model of rational action and decision making to support analysing dialogue. The model approximates principles of behaviour from game theory, and its proof theory makes Gricean principles of cooperativity derivable when the agents' preferences align.

1 Introduction

Grice (1975) and Neo-Griceans model the link between dialogue processing and general principles of rational behaviour by assuming that agents abide by a strong cooperativity principle—namely, people normally believe what they say and help other agents achieve the goals that they reveal through their utterances. This principle provides cooperativity on at least two levels: a basic level that ensures coordination on the conventions governing linguistic meaning (basic cooperativity); and a level concerning shared attitudes towards what is said, including shared intentions (content cooperativity). But not all conversations are content cooperative. For example, Tomm and Dave don't share intentions in (1), taken from chat recordings of an online version of the game *Settlers of Catan* where players negotiate over restricted resources:

(1) a. Tomm: Got any clay to trade for sheep/wheat?
b. Dave: Only got 1 and I'm holding on to it, sorry.

However, even though *Settlers* is a win-lose game where players' interests are often opposed, its players often do share intentions, as they must cooperate to bargain for resources they need in the game.

(2) a. William: can i get a sheep or a wheat?
b. i have too much wood.
c. Cat: i can give you a wheat.
d. William: good
[they exchange 1 wheat for 1 wood]

Conversely, dialogue (3) is content cooperative (and basic cooperative) on the assumption that A and B are constructing a plan to achieve the same goal—that they both eat at Chop Chop:

(3) a. A: Let's go to Chop Chop by car.
b. B: But there's no parking.
c. A: Then let's take the bus.

But (3b) implicates that B *rejects* the intention underlying (3a)—to go to Chop Chop by car. The grounds for the rejecting moves in (1) vs. (3) are different. In (1) Dave can fulfil Tomm's intention but chooses not to, presumably because of his conflicting preferences. In (3) A and B (transiently) have different intentions because of their conflicting beliefs about the optimal way to achieve a shared preference: to eat at Chop Chop.

There have been several attempts to make models of conversation that abide by Gricean principles of cooperativity formally precise; for instance, by expressing axioms in a logic that supports defeasible reasoning about the cognitive states of dialogue agents (e.g., Schulz (2007), Asher and Lascarides (2003)). Such models include (default) axioms Sincerity (following Grice's maxim of Quality (Grice, 1975, p46)): if a cooperative agent says something then he normally believes it. They also include axioms at the level of intentions. For instance, the following default axiom of Strong Cooperativity is adapted from Grice's analysis of what an utterance *means* in cooperative conversation (Grice, 1969, p151): if an agent says something that implies he has a particular intention ϕ that he also intends should be recognised, then a cooperative agent should normally adopt that intention ϕ too (e.g., (Grosz and Sidner, 1990, p430), Asher (in press)). But such formalisations are incomplete because they do not handle cases like (1) where content cooperativity breaks down: the default axioms just given don't apply in such contexts. Nor do they predict when adopting shared intentions is optimal in a strategic setting (e.g., (2)). Furthermore, (3) shows that, rather than expressing content cooperative behaviour in terms of the default adoption of the other agents' *intentions*, it would be better to define it in terms of shared *preferences*; that way, rejection is rational when conflicting beliefs yield a different optimal way to achieve a shared preference.

This paper provides a cognitive model within which one can explore reasoning about the mental states of dialogue agents. We will derive Gricean principles of cooperativity, formalised as defeasible principles, from a characterisation of certain games, using game theory as the foundation of strategic reasoning and also as the basis for linking inferences about conversation to rational action. In game theory, agents act so as to *maximise their expected utility*—*utility* being a measure of preference, and the term *expected* ensuring that decisions about action are made relative to one's beliefs about what the outcomes of the actions will be, including beliefs about what other agents will do. Nevertheless, we argue in Section 2 that game theory on its own provides an incomplete picture, which makes it difficult to use to derive defeasible principles. The symbolic cognitive model presented in Section 3 addresses this problem. It provides axioms approximating rational behaviour from game theory that link dialogue actions and mental states, and its *proof theory* allows us to derive Gricean principles of cooperativity when the agents' preferences align. We relate this approach to prior work in Section 4 and point to future work in Section 5.

2 Our Model

Our model for strategic agents is one that is based on logic and on game theory. Like many others, we use a variant of a Belief Desire Intention (BDI) logic to formalise Gricean implicatures. But because we countenance misdirection and deception as features of strategic conversation, we draw a distinction between **Public and Private** attitudes and thus introduce a new attitude for public commitment. Speaking makes an agent *publicly commit* to some content (Hamblin, 1987). Traditional mentalist models of dialogue, couched within BDI logics, *equate* dialogue interpretation with updating mental states: e.g., interpreting an assertion that p is equivalent to updating one's model of the speaker's mental state to include a belief in p (e.g., Grosz and Sidner (1990)). But they are not equivalent in (4):

(4) a. Loreleil292: Can anyone give me some clay for some wheat?
    b. AMI123: Sorry have none of that!
    [in fact, she has 2 clay]

We interpret (4b) as a negative answer *even if* we know AMI121's beliefs are inconsistent with this. To do justice to this, we follow Asher and Lascarides (2003) and separate the representation and logic of dialogue content from that of cognitive states and then link them via defeasible transfer principles. This separation was originally motivated by calculable implicatures being unavailable as antecedents to surface anaphora; insincerity provides a new motivation.

In common with game theoretic models of conversation (e.g., Parikh (2001)), we adopt a second principle: people say things that will maximise their expected utility. So if the Gricean maxims of conversation hold, they do so because they maximise the agents' expected utility. We also maintain that agents' preferences evolve as dialogue proceeds, at least partly because agents learn about other agents' preferences from what they say and then adjust at least some of their preferences in the light of this information. People's preferences are typically **partial** and get more specific or evolve as they learn more through conversation.

This last assumption poses a problem for orthodox game theory. Game theory assumes each player has a completely defined preference function over the possible actions in the game. It models uncertain and partial information that one player has about another player's preferences and the actions that other play is contemplating performing by a probability distribution over player types, where each type is associated with a complete set of actions and a complete utility function.[1] Game theory, however, does not provide general principles for restricting the set of player types one needs to consider or the probability distributions over them. This gap has bite in modelling conversation, because the possible signals that grammars of natural languages allow are unbounded, as are the coherent signals in context. So dialogue agents generally face the task of isolating their game problem to a set of signals that is small enough to effectively perform inference over, but large enough to yield reliable decisions about optimal actions.

To represent dialogue processing it would be better not to remain silent on how one identifies which player types—and hence which actions and preferences—are relevant, but rather to consider a partial theory or description of the agent's preferences that is updated or revised as one learns more about the agent or one considers actions that one didn't consider before. This is what we do here. This approach yields a more compact and tractable cognitive model and a proof theory in which we can reason, in the light of new evidence, about what type of player to consider in our reasoning. Standard game-theory provides models that can verify the soundness of our proof theoretic reasoning. Whether the sentences in this theory are assigned probabilities is not terribly relevant. But what *is* important is that elements of this theory get revised in the light of new evidence, as Alchourrón et al. (1985) suggest. This can either be done by conditionalising a probability distribution over new evidence, or more symbolically via a theory that incorporates general but defeasible principles about human action and the preferences that underlie them.

In our symbolic model, instead of a probability distribution over every possible complete model of the game we begin with just one partial model. We will describe this model in a way that meshes easily with inferring preferences from observing what agents do. We demonstrate one advantage of this approach here: the proof theory afforded by our symbolic axioms of rational behaviour, which approximate those from game theory, is sufficient to *derive* Gricean principles of cooperativity, among them the default axioms of Sincerity and Strong Cooperativity that we discussed in Section 1. We thus gain a logical link between strategic conversation and content cooperative conversation.

3 Cognitive Modelling

To reason about an agent's motives and actions we use a familiar modal logic: $\mathcal{B}_a\phi$ means agent a believes ϕ, and $\mathcal{I}_a\phi$ means a intends to bring about a state that entails ϕ. We assume that $\mathcal{B}_a$ abides by the modal axioms KD45 (so its accessibility relation in the model is transitive, euclidean and serial); so an agent's beliefs are mutually consistent with one another and closed under logical consequence, and agents have total introspection on their beliefs or lack thereof. We make $\mathcal{I}_a$ abide by the modal axiom D (so its accessibility relation in the model is serial), so contradictory intentions are ruled out. We also assume that intentions are doxastically transparent: i.e., $\mathcal{I}_a\phi \leftrightarrow \mathcal{B}_a\mathcal{I}_a\phi$ is an axiom. We also need a modal operator for public commitment, which is distinct from belief: $P_{a,D}\phi$ means agent a publicly commits to ϕ to the group of agents D. Following Asher and Lascarides (2008), we make $\mathcal{P}_{a,D}$ K45 (one commits to all the consequences of one's commitments and one has total introspection on commitments or lack of them). Unlike belief, commitments can be contradictory because one can declare anything.

[1] Game theory allows players to have imperfect knowledge of what action other players play, but that is not relevant here.

Reasoning about mental states is inherently defeasible, so we add to our logic the weak conditional from Commonsense Entailment (Asher, 1995): $A > B$ means *If A then normally B*. We call this language CL (standing for *cognitive logic*). This logic has many nice properties; for instance, soundness completeness and decidability (Asher, 1995). Decidability is maintained even in a dynamic version of the CL (Asher and Lascarides, 2011), but for the sake of simplicity we will consider the static version here.

The dialogue's logical form creates public commitments in CL: if the logical form stipulates that agent a is committed to K at turn n—so K is the content of a clause or of coherently related dialogue segments—then this makes $\mathcal{P}_{a,d}\mathcal{K}$ true in CL, where $\mathcal{K}$ is the CL-representation of the formula K in the separate logic of dialogue content, and D is the set of dialogue agents (how K maps to $\mathcal{K}$ is detailed in (Lascarides and Asher, 2009) but doesn't concern us here).[2]

3.1 Preferences

Besides a representation of dialogue content and BDI attitudes, we need a symbolic way of representing preferences and commitments to preferences. CP-nets (Boutilier et al., 2004) provide a useful formalism for extracting commitments to preferences from utterances (Cadilhac et al., 2011). Standard CP-nets capture *complete* information: they are a compact representation of a preference order over all the possible outcomes of actions that agents can perform. To represent partial preferences, we build a *partial description of a CP-net* (Cadilhac et al., 2011), which approximates preferences as revealed by dialogue moves. This avoids having to postulate a range of player types, each associated with complete preferences. Instead, agents will reason with and revise partial descriptions of preferences as they observe new evidence through dialogue moves.

A CP-net for an individual agent has two components: a directed *conditional preference graph* (CPG), which defines for each feature F its set of parent features $Pa(F)$ that affect the agent's preferences among the various values of F; and a *conditional preference table* (CPT), which specifies the agent's preferences over F's values for every combination of values in $Pa(F)$ (thus CP-nets have a similar structure to Bayesian belief networks (Pearl, 1988)). The CP-net for a *game* consists of a CP-net for each player. For example, the CP-net in Figure 1 represents a game where A chooses what A and B will eat, and B chooses what they will drink (they must eat and drink the same thing). Agent A's preferred *Food* is fish, but the *Wine* he prefers is dependent on the food: white wine for fish and red for meat. Agent B is indifferent about what he eats, but like A his choice of *Wine* is dependent on what he eats. The logic of CP-nets follows two ranked principles when generating the preference order over every outcome from this compact representation: first, one prefers values that violate as few conditional preferences as possible; and second, violating a (conditional) preference on a parent feature is worse than violating the preference on a daughter feature. So Figure 1 yields the following partial order over all outcomes for each agent:

$$
\begin{tikzcd}
\textit{Food} \arrow[d, bend right, "A"'] \arrow[d, bend left, "B"] \\
\textit{Drink}
\end{tikzcd}
$$

The CPG

Preferences for A:

	fish $\succ_A$ *meat*
fish:	*white* $\succ_A$ *red*
meat:	*red* $\succ_A$ *white*

Preferences for B:

	fish $\sim_B$ *meat*
fish:	*white* $\succ_B$ *red*
meat:	*red* $\succ_B$ *white*

The CPTs

Figure 1: A CP-net for the food and drink game.

(5)
$$
\begin{aligned}
&(\textit{fish}, \textit{white}) \succ_A (\textit{fish}, \textit{red}) \succ_A \\
&\qquad (\textit{meat}, \textit{red}) \succ_A (\textit{meat}, \textit{white}) \\
&\{(\textit{fish}, \textit{white}), (\textit{meat}, \textit{red})\} \succ_B \\
&\qquad \{(\textit{fish}, \textit{red}), (\textit{meat}, \textit{white})\}
\end{aligned}
$$

There are efficient algorithms for identifying the (unique) optimal strategy in this case (e.g., Bonzon (2007)): i.e., to eat fish and drink white wine.

Dialogue interpretation yields commitments to preferences that are *partial*. For example, by

[2] K captures *all* a's current commitments, including ongoing commitments from prior turns. So there is no need to conjoin an agent's commitments from each turn in CL.

$$
\begin{array}{lll}
\textit{Receive} & & \\
\downarrow C & & \textit{r-wood} \succ_C \neg\textit{r-wood} \\
\textit{Give} & \textit{r-wood}: & \textit{g-wheat} \succ_C \textit{g-sheep}
\end{array}
$$

Figure 2: Cat's commitments to preferences in dialogue (2). *Receive* and *Give*'s values are *r-x* and *g-x* where x is *wheat*, *clay*, *sheep*, *rock* or *ore*.

uttering (2c) Cat commits to the partial CP-net in Figure 2: In words, Cat would rather receive wood than not, and given her preference for receiving wood she would rather give wheat than sheep. This is computed recursively from the discourse structure of Cat's commitment (Cadilhac et al., 2011), although we don't detail the mapping here.[3]

Crucially, the CP-net description in Figure 2 is partial: it doesn't reveal Cat's preferences for giving wheat or sheep in a context where she doesn't get wood—(2c) says nothing about that. Cat's actual preferences may also differ from these commitments (e.g., because of insincerity). So to choose an optimal action, agents face a complex calculation in (defeasibly) estimating an agent's complete actual preferences from commitments to them.

Accordingly, we treat the preference statements in the CP-net descriptions as formulas within a background theory that provides defeasible inferences about preferences and behaviour. Our background theory is CL; so CL must be able to express and reason about descriptions of CP-nets. Specifically, we complete the partial information in a CP-net description by adding assumed preferences that defeasibly follow from it via the default axioms in CL, with agents defaulting to being indifferent among values for features for which preference information is missing entirely. In logical terms this means we will have formulae in CL of the form: $\chi > (\phi\colon \psi \succ_a \neg\psi)$, where χ is a well-formed formula of CL.[4] In other words, if χ is true then normally the description of a's preferences includes $\phi\colon \psi \succ_a \neg\psi$ (note that the antecedent χ may express information about preferences too). We'll give some examples of such formulae in the next section. Further, CL's nonmonotonic inferences about an agent's preferences may change if the range of actions that are considered to be a part of the game changes (though we forego specific examples here). Overall, through the (nonmonotonic) logic of CL's $>$, one can support decisions about what action to perform even if knowledge of preferences is partial.

CL can now link preferences to other propositional attitudes. Indeed, choosing optimal actions requires a link between preference and *belief*: since a (joint) CP-net G can include variables whose values one doesn't control, one needs to check that one's optimal state(s) are not doxastically improbable (this is a crude way of ensuring that agents act so as to maximise *expected* utility rather than acting with wishful thinking about what's feasible). We supply a notion of doxastic improbability in CL via its nonmonotonic consequence relation: i.e., a state is *belief compliant* if its negation does not defeasible follow from the premises and background theory of CL axioms. So to identify an agent's optimal belief-compliant state(s), we filter out any optimal state that is defeasibly inconsistent with his beliefs (as we mentioned in Section 3, this is decidable). Within CL this leads to the definition of a *CP-solution*$_a(\phi, G)$ for agent a and (joint) CP-net G:

Definition 1 CP-solution$_a(\phi, G)$ *holds iff:*

1. *a is a player in the joint CP-net G; and*
2. *$s \vdash \phi$ for every belief-compliant optimal state s of G. I.e., where Γ is the premises—in other words,* CL*'s background theory plus information about the mental states of the players in G—we have $\Gamma \not\mathrel{|\!\sim} \mathcal{B}_a \neg s$, and for*

[3] We take William's and Cat's commitments in (2) to be as follows (Lascarides and Asher, 2009). William's turn commits him to *Plan-Elab*(a, b), which means that he commits to the contents of both (2a) and (2b) and to (2b) elaborating a plan to achieve the goal underlying (2a) (that goal is to obtain a sheep or wheat, and the plan afforded by (2b) is to get one of these by trading wood). Cat's utterance (2c) commits her to *Plan-Elab*(π, c), where π is William's first turn (with content *Plan-Elab*(a, b)). Cadilhac et al's (2011) recursive algorithm yields Figure 2 from *Plan-Elab*(π, c).

[4] Since the features in our CP-nets all take finite values, they can be represented in CL using Boolean variables.

any state s' that is strictly more optimal in G than s, $\Gamma \mid\!\sim \mathcal{B}_a \neg s'$ holds.

For example, if B's model of A's and his own preferences are those in Figure 1, then by Definition 1 *CP-solution*$_B$(*fish*$\wedge$*white*$, G)$ holds: while *meat* $\wedge$ *red* is equally preferred by B, it is not belief compliant because G defeasibily entails that A will choose *white* and not *red*. We'll see this in the next section, when we use CP-solutions to define CL axioms that approximate principles of rational action from game theory.

3.2 Axioms of Rationality

To encode means-end reasoning of rational agents in our symbolic model, we need CL axioms that make agents *pay-off maximisers* (cf. rationality from game theory) and *basic cooperative*. Pay-off maximisers intend actions that are an optimal trade-off between their preferences and their beliefs about what's possible; and an agent intending ψ means in the context of his current beliefs he prefers ψ to all alternative actions. We capture these two principles with the axioms **Maximising Utility** (a) and (b):

Maximising Utility:

a. $(G \wedge \textit{CP-solution}_a(\psi, G)) > \mathcal{I}_a\psi$
b. $(\mathcal{I}_a\psi \wedge \textit{player}(i, G)) > \textit{CP-solution}_a(\psi, G)$

Maximising Utility part (a) ensures a intends ψ only if ψ follows from all belief-compliant optimal states (by Definition 1). Indeed, agent a's intentions are conditional on *all* of a's beliefs (thanks to Definition 1) and *all* of a's preferences and those of any player that affect a's preferences. The latter property follows because the weak conditional $>$ validates the Penguin Principle—i.e., default consequences of rules with more specific antecedents override conflicting defaults from less specific antecedents. So if a more specific game G' is known to hold and it yields conflicting intentions to those resulting from G, then the intentions from G' are inferred and those from G aren't. Axiom (b) likewise conditions a's preference for ψ on all his beliefs (thanks to Definition 1). It yields (default) constraints on G from intentions: if one knows $\mathcal{I}_a\psi$ and nothing about G or about a's beliefs, then the minimal CP-net G that satisfies the default consequence is simply the global preference $\psi \succ_a \neg\psi$. As agents converse, each dialogue action may reveal new information about intentions, and via Maximising Utility part (b) this imposes new constraints on G. But while Maximise Utility part (b) is conservative about exactly which of a's beliefs his preference for ψ is conditioned on, his dialogue moves can reveal more precise information—e.g., the utterance *I want to go to the mall to eat* should be sufficient to infer *eat* : *mall* $\succ_i$ ¬*mall*. A detailed algorithm for extracting preferences and dependencies among them from conversation is detailed in Cadilhac et al. (2011), but the details of this aren't relevant for our purposes here.

Basic cooperativity follows from an axiom that makes all agents intend that their commitments be shared among all the other dialogue agents:

Intent to Share Commitment:
$(b \in D \wedge \mathcal{P}_{a,D}\phi \wedge \neg\mathcal{P}_{b,D}\phi) > \mathcal{P}_{a,D}\mathcal{I}_a\mathcal{P}_{b,D}\phi$

If a commits, when addressing b (among others), to content ϕ and b hasn't committed to this yet, then normally a is also committed to intending that b so commit. This rule captures basic cooperativity because b committing to a's commitments entails he *understands* a's commitments (Clark, 1996). Indeed, it captures something much stronger than basic cooperativity—an intention that your contribution be *accepted* by others. While this is stronger than basic cooperativity, we think it's *rational* even in non-cooperative dialogue contexts: why commit to content if you don't intend that others accept the commitment? In addition, we regiment a constraint on assertions proposed by (Perrault, 1990, p180), by refining this axiom for assertions: when a's address to b commits him to an assertion $\mathcal{K}$, then normally $\mathcal{P}_{a,D}\mathcal{I}_a\mathcal{B}_b\mathcal{K}$.

Now let's examine more carefully the special case of Gricean cooperativity. We start by defining a Grice Cooperative game:

Definition 2 *A game is* **Grice Cooperative (GC)** *just in case for any of its players a and b*

1. their speech acts normally have their con-

ventional purpose (e.g., they normally ask a question so as to know a true answer); and

2. $(\phi : \psi \succ_a \neg\psi) > (\phi : \psi \succ_b \neg\psi)$ *(i.e., the agents' preferences normally align).*

We can now prove all the axioms in Fact 1.

Fact 1 Sincerity: $(\mathcal{P}_{a,D}\phi \wedge GC) > \mathcal{B}_a\phi$
Sincerity for Intentions:
$(\mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC) > \mathcal{I}_a\phi$
Sincerity for Preferences:
$(\mathcal{P}_{a,D}(\phi : \psi \succ_a \neg\psi) \wedge GC) > \phi : \psi \succ_a \neg\psi$
Competence:
$(\mathcal{P}_{a,D}\phi \wedge \mathcal{P}_{b,D}?\phi \wedge a, b \in D) \rightarrow ((\mathcal{B}_b\mathcal{B}_a\phi \wedge GC) > \mathcal{B}_b\phi)$
Cooperativity:
$(b \in D \wedge \mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC) > \mathcal{I}_b\phi$

These axioms make any *declared* belief, intention or preference in a GC conversation normally an *actual* belief, intention or preference too (cf. the Gricean maxim of Quality (Grice, 1975, p45)). Competence makes belief transfer the norm (if b asked whether ϕ). This default likewise follows from Grice's Maxim of Quality as he described it in (Grice, 1989, p371): he stipulates that in order to contribute to a conversation via the Maxim of Quality, one must say what is true. To do otherwise is not to contribute inferior information; rather, it contributes no information at all. Furthermore, Lewis (1969) argues persuasively that unless such a principle of competence forms the basis of cognitive modelling, then one cannot construct a sound philosophical argument that explains why linguistic conventions come into being in the first place, or why we assume that a speaker whom we understand is speaking the same language as we are—a hallmark of basic cooperativity. Finally, Cooperativity makes a declared individual intention normally a shared actual intention (recall the Gricean notion of utterance meaning in conversation (Grice, 1969, p151) and the corresponding notion of Strong Cooperativity from Section 1). Such principles of sincerity and cooperativity are usually taken as primitive axioms in BDI approaches to dialogue; here, we *derive* them when agents D are players in a joint game G that satisfies Definition 2.

Outline Proofs: Sincerity: Suppose $\mathcal{P}_{a,D}\phi$ and GC hold and moreover that ϕ expresses a proposition that is capable of being believed. Then we'll show that if all the normal GC consequences hold (see Definition 2), then $\mathcal{B}_a\phi$ must also hold.

By Intent to Share Commitment, $\mathcal{P}_{a,D}\mathcal{I}_a\mathcal{B}_b\phi$ defeasibly follows from our premises for any $b \in D$. By Maximising Utility and the fact that $\mathcal{I}$ is a D modality, $\mathcal{I}_a\mathcal{B}_b\phi$ defeasibly implies $\mathcal{B}_b\phi \succ_a \neg\mathcal{B}_b\phi$. Upon learning of a's commitment and the fact that the game is GC (in particular, clause 1 of Definition 2 means that the preference underlying a's move ϕ that we have just derived is a's actual preference), we infer $\mathcal{B}_b\phi \succ_b \neg\mathcal{B}_b\phi$. Assume further that belief preferences pattern after factual preferences. That is:

$$(\mathcal{B}_b\phi \succ_b \neg\mathcal{B}\phi) \rightarrow (\phi \succ_b \neg\phi)$$
$$(\neg\mathcal{B}_b\phi \succ_b \mathcal{B}\phi) \rightarrow \neg(\phi \succ_b \neg\phi)$$

So $\phi \succ_b \neg\phi$. Now suppose that $\neg\mathcal{B}_a\phi$. Then assuming we prefer our belief actions when we have them, $\neg\mathcal{B}_a\phi \succ_a \mathcal{B}_a\phi$, and therefore $\neg(\phi \succ_a \neg\phi)$. Thus the game cannot be a normal GC game, contrary to our assumptions. So $\mathcal{B}_a\phi$. Now, **Weak Deduction** is a valid rule of the weak conditional $>$ (Asher, 1995): if $\Gamma, \phi \mid\!\sim \psi$, $\Gamma \not\mid\!\sim \psi$ and $\Gamma \not\mid\!\sim \neg(\phi > \psi)$ then $\Gamma \mid\!\sim (\phi > \psi)$. So Weak Deduction yields the desired $>$ statement, $(\mathcal{P}_{a,D}\phi \wedge GC) > \mathcal{B}_a\phi$. $\square$. We can also derive (though we don't show it here) a stronger version of Sincerity where a doesn't believe alternatives to what he said, yielding scalar implicatures.

Sincerity for Intentions: Suppose $\mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC$. By Sincerity (which we've just proved), $\mathcal{B}_a\mathcal{I}_a\phi$. Since intentions are doxastically transparent (i.e. $\mathcal{B}_a\mathcal{I}_a\phi \leftrightarrow \mathcal{I}_a\phi$), the result follows with an application of Weak Deduction. $\square$.

Sincerity for Preferences is proved in a similar way, using also the assumption that preferences are doxastically transparent. $\square$.

Competence: Suppose $\mathcal{P}_{b,D}?\phi \wedge \mathcal{P}_{a,D}\phi \wedge b \in D \wedge \mathcal{B}_b\mathcal{B}_a\phi$ and a GC game. Given Definition 2, the intention that normally underlies asking a question (i.e., to know an answer) and Maximising Utility ensures that b's asking ϕ implies $\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi \;\succ_b\; \neg(\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi)$. So by GC (i.e., the agents' preferences normally align), we also have: $\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi \succ_a \neg(\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi)$. By Maximising Utility we can assume that b's asking a question together with a's response are both optimal moves in equilibrium. These moves then should realise the preference $\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi \succ_b \neg(\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi)$. Furthermore, by Sincerity, $\mathcal{B}_a\phi$. There are two choices now: either a is trustworthy or not. If a is not trustworthy, then his commitment to ϕ is no indication of its truth. But then there is a move (do not ask a whether ϕ) that would have been more advantageous for b (given that listening to someone and processing the response is a cost). So given that $\mathcal{P}_{b,D}?\phi$ is the equilibrium move—in other words, this is a move that is optimal for b in that it maximises his expected utility—b must believe a to be trustworthy, and so $\mathcal{B}_b\phi$. Using Weak Deduction thus yields Competence. $\Box$.

Cooperativity: Assume $b \in D \wedge \mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC$. By Sincerity for Intentions, we have $\mathcal{I}_a\phi$. By Maximising Utility, we can infer *CP-solution*$_a(\phi, G)$, where G is the GC game with at least a and b as players. By GC and Competence, this defeasibly entails *CP-solution*$_b(\phi, G)$. And so Maximising Utility yields $\mathcal{I}_b\phi$. Using Weak Deduction gets us the desired $>$ statement. $\Box$.

Intention and belief transfer in a GC conversation is a *default*: even if preferences align, conflicting beliefs may mean agents have different CP-solutions making their intentions different too (by Maximising Utility), and Competence may apply but its consequent isn't inferred. Thus rejection and denial occur in GC dialogues (see (3)). On the other hand, in GC environments interpretations are normally credible: e.g., by Sincerity and Competence B's assertion (3b) yields belief transfer that there's no parking. This is a simple, symbolic counterpart to the much more elaborate result concerning credibility from Crawford and Sobel (1982).

4 Related Work

In contrast to Gricean formalisations in BDI logics, we have conditioned Gricean behaviour on shared *preferences* rather than shared *intentions* (see Definition 2) and we have derived Gricean axioms from a more general axiomatisation of human behaviour rather than treating them as primitive.

Signalling games provide a basis for predicting conversational implicatures (e.g., Parikh (2001), van Rooij (2004)) and also insincerity—the less aligned the preferences, the less credible the signals (Crawford and Sobel, 1982). But signalling models either take a signal to mean whatever it is optimal for it to mean (thereby bypassing linguistic convention) or the mapping $[\![.]\!]$ from signals to meaning is fixed and monotonic (e.g., Farrell (1993), Franke (2010)), with pragmatic interpretations being entirely epistemic in nature: they arise when the optimal interpretation of s is distinct from $[\![s]\!]$. Our model differs in its view of *conventional meaning*: while we acknowledge that some pragmatic inferences are epistemic (e.g., see **Sincerity**), we also believe that $[\![s]\!]$ goes beyond lexical and compositional semantics because it is constrained to be *coherent* (Lascarides and Asher, 2009). But this makes computing $[\![s]\!]$ defeasible, which reflects the fact that all inferences about coherence are defeasible. So in non-cooperative conversation, an interlocutor must test rigorously his defeasible inference about what the speaker is publicly committed to, as well as test the credibility of that commitment (i.e., whether the speaker believes it). We hope that CL can model such tests, but leave this to future work.

5 Conclusions

We have proposed a qualitative model of cognitive reasoning with several desirable features for modelling dialogue: it supports reasoning with partial information about preferences; and it distinguishes the public commitments one makes through utterances and private mental states that affect and are affected by them. The axioms of the cognitive logic approximate rational action from game theory and compel agents to be basic cooperative. We showed that Gricean principles of sincerity and cooperativity are derivable from them when the agents' preferences normally align.

We have focused here entirely on the cognitive model; linking it to dialogue content is ongoing work. The cognitive logic should also be dynamic since dialogue actions trigger changes to mental states: our static CL can be made dynamic with no cost to complexity by exploiting public announcement logic (Asher and Lascarides, 2011). Finally, progress in analysing strategic conversation requires an extensive study of data in many domains: e.g., political debate, commercial negotiations, courtroom cross examination and others. The *Settlers* dialogues cited here are all taken from our ongoing corpus collection effort, in which utterances are aligned with machine readable game states. We hope to release this corpus, labelled with rich semantic and cognitive information, in due course.

Acknowledgements: This work is supported by ERC grant 269427 (STAC).

References

CE Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meeting contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

N. Asher. Commonsense entailment. In G. Crocco, L. Farinas, and A. Herzig, editors, *Conditionals: From Philosophy to Computer Science*, pages 103–145. OUP, 1995.

N. Asher. Implicatures in discourse. to appear in *Lingua*, in press.

N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

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

N. Asher and A. Lascarides. Reasoning dynamically about what one says. *Synthese*, 183(1):5–31, 2011.

E. Bonzon. *Modélisation des Interactions entre Agents Rationnels: les Jeux Booléens*. PhD thesis, Université Paul Sabatier, Toulouse, 2007.

C. Boutilier, R.I. Brafman, C. Domshlak, H.H. Hoos, and David Poole. Cp-nets: A tool for representing and reasoning with conditional *ceteris paribus* preference statements. *Journal of Artificial Intelligence Research*, 21:135–191, 2004.

A. Cadilhac, N. Asher, F. Benamara, and A. Lascarides. Commitments to preferences in dialogue. In *Proceedings of SIGDIAL*, pages 204–215, 2011.

H. Clark. *Using Language*. Cambridge University Press, Cambridge, England, 1996.

V. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

J. Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behaviour*, 5:514–531, 1993.

M. Franke. Semantic meaning and pragmatic inference in non-cooperative conversation. In T. Icard and R. Muskens, editors, *Interfaces: Explorations in Logic, Language and Computation*, pages 13–24. Springer-Verlag, 2010.

H.P. Grice. Utterer's meaning and intentions. *Philosophical Review*, 68(2):147–177, 1969.

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

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

B. Grosz and C. Sidner. Plans for discourse. In J. Morgan P. R. Cohen and M. Pollack, editors, *Intentions in Communication*, pages 417–444. MIT Press, 1990.

C. Hamblin. *Imperatives*. Blackwells, 1987.

A. Lascarides and N. Asher. Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26 (2):109–158, 2009.

D. Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

P. Parikh. *The Use of Language*. CSLI Publications, Stanford, California, 2001.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffmann, 1988.

R. Perrault. An application of default logic to speech act theory. In J. Morgan P. R. Cohen and M. Pollack, editors, *Intentions in Communication*, pages 161–186. MIT Press, 1990.

K. Schulz. *Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals*. PhD thesis, University of Amsterdam, 2007.

R. van Rooij. Signalling games select horn strategies. *Linguistics and Philosophy*, 27:493–527, 2004.

Meanings as proposals:
a new semantic foundation for Gricean pragmatics

Matthijs Westera
Institute for Logic, Language and Computation
University of Amsterdam
m.westera@uva.nl

Abstract

A disjunction may pragmatically imply that only one of the disjuncts is true. The traditional Gricean account of this exhaustivity implicature is not without problems. Nevertheless, we think that not the Gricean picture itself, but the underlying conception of meanings as chunks of information may be unfit. Starting instead from a conception of meanings as proposals, within the framework of inquisitive semantics, we develop, algebraically characterise and conceptually motivate a formal semantics and pragmatics, the latter still Gricean in spirit.

Among the difficulties we discuss and resolve are the problem of characterising relevant alternatives, the problem of embedded implicatures and the unwanted negation problem. The analysis is extended to a pragmatic account of mention-some questions.

1 Introduction[1]

1.1 Some problems for a Gricean pragmatics

Sentence (1) asserts that it is rainy or windy, and may pragmatically imply that not both are the case.

(1) It's rainy or windy.

The existence of a 'not both'-implicature is suggested by the fact that one can say 'it's rainy or windy or both' without a sense of redundancy, as well as the fact that a 'both'-response to (1) would be slightly unexpected.

If we translate (1) as $p \vee q$, the traditional Gricean account of the 'not both'-implicature of (1) reads as follows (roughly adopted from (Chierchia, Fox, & Spector, 2008)):

Reasoning pattern 1 (Traditional account)

1. *The initiator said $p \vee q$.*
2. *If $p \vee q$ is relevant, then presumably p, q and $p \wedge q$ are too.*
3. *The initiator could have said $p \wedge q$, which is stronger and relevant.*
4. *The reason for the initiator choosing $p \vee q$ over $p \wedge q$ might be that he does not believe that $p \wedge q$.*
5. *It is likely that the initiator has an opinion as to whether $p \wedge q$ is true.*
6. *Hence, the initiator must believe that $p \wedge q$ is false.*

There are a number of weaknesses in this approach, many of which have to do with **characterising the set of relevant alternatives** used in step 2 of the reasoning above. As argued in (Chierchia et al., 2008), under certain natural assumptions regarding the concept of relevance, the set of relevant alternatives grows too big to yield any implicature at all (except implicatures of ignorance). This has been partially solved by postulating lexically specified *scales* of alternatives (Horn, 1972). However, it is not so clear conceptually and technically where the scales come from, and they are not immune to trouble, either.

[1] Many thanks to Floris Roelofsen and Jeroen Groenendijk and three anonymous reviewers for helpful comments. Any remaining mistakes are of course my own. Financial support from the Netherlands Organization for Scientific Research (NWO) is gratefully acknowledged.

For instance, in what Spector (2007) calls the **unwanted negations problem**, the Gricean approach predicts that a triple disjunction $p \vee q \vee r$, given its scalar alternative $(p \wedge q) \vee r$, would imply $\neg r$ - something which is clearly not the case.

(Chierchia et al., 2008) discuss the problem of **embedded implicatures**, i.e., implicatures that seem to arise from within the scope of, e.g., a quantifier.

(2) a. Each student read *Othello* or *King Lear*.
  b. Each student read *Othello* and *King L.*.

Since (2b) is a scalar alternative to (2a), the above reasoning predicts the implicature that the speaker believes that not every student read both. But this is arguably too weak. What should come out is the implicature that every student did not read both. For this, the alternatives would have to be computed within the scope of the quantifier, but this seems to go against a genuinly Gricean pragmatics.

Another weakness, independent of how the set of alternatives is characterised, is what (Sauerland, 2005) calls the **epistemic step**. Going from step 4 to 6 in reasoning pattern 1 above requires a strengthening from not believing, to believing that not, i.e., from $\neg\Box\varphi$ to $\Box\neg\varphi$. This strengthening does not follow from the Gricean maxims and logic alone, but requires an extra, stipulated assumption, given in step 5 above.

It has been noted that the implicature of (1) is perhaps an instance of a larger class of **exhaustivity** implicatures (Rooij & Schulz, 2006). For example, similar pragmatic strengthening seems to be going on in (3), paraphraseable as 'it's either only rainy, or only windy, or both rainy and windy':

(3) It's rainy or windy or both.

However, it is not clear what, if any, the relevant alternatives to (3) should be that would yield an exhaustivity implicature. Replacing any disjunction(s) in (3) by a conjunction results in a formula equivalent to $p \wedge q$, incorrectly predicting the same 'not both'-implicature as for (1).[2]

Related to the issue of exhaustivity are **mention-some questions** (e.g., Rooij & Schulz, 2006). Such questions do not ask for an exhaustive answer, but rather are satisfied with the responder mentioning some possible instances:

(4) a. A: I will pick up the key this afternoon. Will your father or mother be home?
  B: My father will be home.
  b. A: Where can I buy toilet paper around here?
  B: In the shop around the corner.

Here B's responses do not imply that A's mother will not be home or that the shop around the corner is the only place that sells toilet paper. This lack of exhaustivity can be tied to the pragmatics debate by observing that the indicative counterpart of (4) does not imply exhaustivity:

(5) You can pick up the key this afternoon. My father or mother will be home.

However, reasoning pattern 1 above does not provide any means for canceling the 'not both' implicature in this case.

We note that the authors cited so far have all come up with solutions, partial or whole, to these difficulties for the traditional Gricean account. For reasons of space, however, we will not discuss these solutions in the present paper.

1.2 Aims and approach

So far, we have discussed (all too briefly) several difficulties for a traditional Gricean pragmatics, that have to do with characterising relevant alternatives (the requirement for scales, the unwanted negation problem), the ad-hoc nature of the epistemic step, and a more general account of exhaustivity (example (3), mention-some questions). This paper is devoted to overcoming them while maintaining the Gricean spirit.

Our approach is to base an in essence Gricean account of the implicatures of (1) and (3) upon a new conception of meaning. Existing accounts are built upon a classical, boolean semantics, that models meanings as chunks of information, or upon a dynamic semantics, based on the view of meaning as context change potential (e.g., Rooij & Schulz, 2006). We follow the framework of Inquisitive Semantics in taking this one step further, regarding meaning as *information exchange*

[2] We assume that for the aims and approach of this paper, the two sentences (1) and (3) can be straightforwardly translated into propositional logic as $p \vee q$ and $p \vee q \vee (p \wedge q)$, respectively.

potential (Groenendijk & Roelofsen, 2009; Ciardelli, Groenendijk, & Roelofsen, 2009; Roelofsen, 2011). There are various ways to make this slogan more concrete, and how this is done will determine properties of the resulting semantics and of the pragmatics built upon it.

Basic inquisitive semantics (InqB) follows from a conception of meanings as *requests for information* (Roelofsen, 2011). InqB has the merit that uttering a disjunction introduces several semantic alternatives, among which a responder is offered a choice. This enrichment of the semantics provides new handles for the pragmatics. Indeed, InqB has been used as a semantic foundation for a pragmatic account of the 'not both'-implicature of (1), that avoids the problematic 'epistemic step' described above (Groenendijk & Roelofsen, 2009). However, InqB, and thereby the pragmatics, does not distinguish between (1) and (3), yielding wrong predictions. Section 4 contains a brief comparison of our approach with (Groenendijk & Roelofsen, 2009).

Unrestricted inquisitive semantics (InqU), as defined in (Ciardelli et al., 2009; Ciardelli, 2010), is a more promising candidate for an account of (1) and (3). For one, it shares with InqB the merit that disjunction introduces alternatives. Second, InqU, unlike InqB, assigns distinct meanings to (1) and (3), offering at least a semantic handle for a Gricean account to also differentiate between them. For this reason, we will base our pragmatic account upon InqU.

InqU lacks the thorough conceptual motivation and algebraic characterisation that (Roelofsen, 2011) developed for InqB. Developing a pragmatic account of (1) and (3) based on InqU cannot be achieved (both technically and conceptually) without first filling in some of the gaps in our understanding of InqU. We will do so by motivating a version of InqU from scratch, starting from a particular conception of meaning, and characterising it algebraically.

Based on this semantics, our essentially Gricean account of examples (1) and (3) will turn out to be technically remarkably simple and conceptually illuminating, and it overcomes all of the mentioned weaknesses of the traditional approach.

2 Unrestricted inquisitive semantics

InqU, as defined in (Ciardelli et al., 2009), is based on a view of meanings as *proposals* to update the common ground in one of several ways, or, in the same paper and in the same breath, as proposals to take certain possibilities into consideration, or to draw *attention* to those possibilities. The road from this conceptual stance to the fully-fledged semantics has not been paved, and the endpoint, i.e., the semantics, has not been characterised algebraically. This is what we attempt in the current section.

2.1 Meanings as proposals

We consider only the language of propositional logic:

Definition 1 (Syntax) *For φ ranging over formulae, p over proposition letters:*
$\varphi := p|\bot|(\varphi \vee \varphi)|(\varphi \wedge \varphi)|(\varphi \rightarrow \psi)$,
with $\neg\varphi := \varphi \rightarrow \bot$.

The semantics for this language is defined relative to a suitable model:

Definition 2 (Model) *A model M is a tuple $\langle \mathbf{W}, I \rangle$, where $\mathbf{W}$ is a set of worlds and I is an interpretation function that, relative to a possible world, maps each proposition letter to a truth value.*

Based on a model, an epistemic state is defined as any subset of the set of possible worlds of the model:

Definition 3 (Epistemic state) *An epistemic state based on the model $\langle \mathbf{W}, I \rangle$ is a set $s \subseteq \mathbf{W}$.*

We think of meanings as proposals. One does not propose a piece of information; rather, one proposes *doing something* with that information, such as updating the common ground with it. Hence, we define meanings, *proposals*, as sets of functions on epistemic states:

Definition 4 (Proposal [to be refined]) *A proposal based on the model $\langle \mathbf{W}, I \rangle$ is a set of functions on epistemic states based on $\langle \mathbf{W}, I \rangle$, i.e., functions $f : \wp\mathbf{W} \rightarrow \wp\mathbf{W}$.*

Because in the present paper we will not be concerned with, e.g., revision mechanisms, we restrict ourselves to functions that are *eliminative* and *distributive*. This allows us to simplify the definition of the resulting semantics, and will make it look and feel like InqB, despite the conceptual shift, as well as InqU in (Ciardelli et al., 2009), as we will see shortly. A function on states is eliminative iff it only *eliminates* worlds, i.e., it does not change the worlds or create new worlds. Conceptually, this means that we consider only functions that model information *growth*, not *loss*; i.e., all functions are actual *update* functions.

Definition 5 (Eliminativity) *$f : \wp \mathbf{W} \to \wp \mathbf{W}$ is eliminative iff $\forall s \subseteq \mathbf{W}$, $f(s) \subseteq s$.*

A function is distributive if we could, instead of applying the function to a state s, apply the function to all singleton substates of s, take the union of their outputs, and obtain the same result. In other words, this means that updates operate locally on worlds, not necessarily globally on states.

Definition 6 (Finite distributivity) *$f : \wp \mathbf{W} \to \wp \mathbf{W}$ is finitely distributive iff $\forall s, s' \subseteq \mathbf{W}$, $f(\varnothing) = \varnothing$ and $f(s \cup s') = f(s) \cup f(s')$.*

Any eliminative, distributive function can be fully characterised by its effect on the uninformed state $\mathbf{W}$ (Benthem, 1989):

Fact 1 (Update decomposition) *For all $f : \wp \mathbf{W} \to \wp \mathbf{W}$, if f is eliminative and finitely distributive, then for all $s \subseteq \mathbf{W}$, $f(s) = f(\mathbf{W}) \cap s$.*

This means that every such update function f corresponds to a unique static object $f(\mathbf{W})$. We will call such objects 'updates-as-states', or just 'updates' when no confusion can arise. (We do not call them 'states', because even though that is what they are, it is not what they represent, conceptually.) Using this result, we refine the definition of proposals to be sets of updates-as-states:

Definition 7 (Proposal) *A proposal A based on the model $\langle \mathbf{W}, I \rangle$ is a set of updates-as-states based on $\langle \mathbf{W}, I \rangle$, i.e., $A \subseteq \wp \mathbf{W}$. Let $[\varphi]$ denote the proposal denoted by a formula φ.*

What $[\varphi]$ consists in will be defined by the semantics.

Through Fact 1, a proposal becomes the same kind of object as a proposition in InqB, i.e., a set of states. However, crucially, it *represents* a different kind of object, namely, a set of update functions. Furthermore, we *think* of proposals in a different way. How we think of proposals must be expressed in a meta-language, for which we choose English.

Definition 8 (The Proposal View) *Every formula φ is paraphrasable as 'let's execute one of the updates in $[\varphi]$'.*

This view on meaning will determine how the semantics is to be defined.

2.2 Conjunction and disjunction

Let us first look at the semantic operation that should underly a **conjunction** of sentences. Meanings, spelled out in our meta-language according to the Proposal View, behave as follows under conjunction:

Observation 1 (Behaviour of conjunction) *Let's do one of the updates in A and let's do one of the updates in B ≡ Let's do two updates, one in A and one in B ≡ Let's do one of the updates in $A \sqcap B := \{a \cap b : a \in A, b \in B\}$.*[3]

Hence, we will take the semantics of conjunction to be pointwise intersection.

The proposal $\{\mathbf{W}\}$ is the identity element for pointwise intersection, i.e., for all $A \in \wp\wp\mathbf{W}$, $A \sqcap \{\mathbf{W}\} = A$. Pointwise intersection is associative and commutative. It is not idempotent: if a proposal, consisting of multiple updates, is made twice, a different update can be chosen the first and the second time, and both of them executed, giving a different result than if the proposal had been made only once (cf. footnote 2). These properties imply that the set of proposals with pointwise intersection and its identity element form a *commutative monoid*:

Fact 2 *$\langle \wp\wp \mathbf{W}, \sqcap, \{\mathbf{W}\} \rangle$ is a commutative monoid:*

1. $A \sqcap \{\mathbf{W}\} = A$
2. $A \sqcap (B \sqcap C) = (A \sqcap B) \sqcap C$
3. $A \sqcap B = B \sqcap A$

Because pointwise intersection is not idempotent, it cannot give the meet with respect to any partial order (the non-idempotency would be in conflict with the reflexivity of the order). However, commutative monoids come with a partial order, called the *divisibility order*, with respect to which pointwise

[3] If A and B are the same proposal, it is not evident that pointwise intersection is indeed adequate. For instance, 'let's have coffee or tea, and let's have coffee or tea' would be equivalent to 'let's have coffee, tea *or both*'. However, a dynamic stance on conjunction ('and then') makes this result acceptable.

intersection *would have* given the meet, had it been idempotent.

Definition 9 (Divisibility order)
$A \leq_{\sqcap} B$ *iff* $\exists C. B \sqcap C = A$.

This can be read as follows: $A \leq_{\sqcap} B$ iff A can be $\sqcap$-decomposed, i.e., factorized, into B and some other proposal C, i.e., iff B is a divisor of A.

Let us now turn to the semantic operation that corresponds to **disjunction**. We spell out the Proposal View again:

Observation 2 (Behaviour of disjunction) *Let's do one of the updates in A or let's do one of the updates in B ≡ Let's do one of the updates in A or one of the updates in B ≡ Let's do one of the updates in $A \cup B$.*

Hence, we will take take the semantics of disjunction to be given by set union.

The proposal $\varnothing$ is the identity element for union, and union is associative, commutative and idempotent, so we have:

Fact 3 $\langle \wp\wp \boldsymbol{W}, \cup, \varnothing \rangle$ *is a commutative, idempotent monoid, i.e.:*

1. $A \cup \varnothing = A$
2. $A \cup (B \cup C) = (A \cup B) \cup C$
3. $A \cup B = B \cup A$
4. $A \cup A = A$

Every commutative, idempotent monoid has a partial order with respect to which it is a join-semilattice, and the operation a join operator. We will call this the *semilattice order*. It can be defined analogously to the divisibility order, but with $\cup$ instead of $\sqcap$, but happens to correspond to the inverse of set inclusion.

Definition 10 (Semilattice order) $A \geq_{\cup} B$ *iff* $\exists C. B \cup C = A$ *(iff $A \cup B = A$ iff $B \subseteq A$).*

Fact 4 $\langle \wp\wp \boldsymbol{W}, \geq_{\cup} \rangle$ *is a join-semilattice, with $\cup$ as join.*

Union and pointwise intersection interact in the following ways. First, $\varnothing$, the identity element for $\cup$, is an annihilator for $\sqcap$, i.e., $\varnothing \sqcap A = A \sqcap \varnothing = \varnothing$. Second, $\sqcap$ distributes over $\cup$. These properties imply that the two monoids together form a *commutative, idempotent semiring*, i.e., a semiring with the unional properties that the first operation ($\cup$) is idempotent and the second operation ($\sqcap$) commutative.

Fact 5 (Algebraic characterisation)
$\langle \wp\wp \boldsymbol{W}, \cup, \sqcap, \varnothing, \{\boldsymbol{W}\} \rangle$ *is an idempotent semiring, i.e.:*

1. $\langle \wp\wp \boldsymbol{W}, \cup, \varnothing \rangle$ *is a commutative, idempotent monoid;*
2. $\langle \wp\wp \boldsymbol{W}, \sqcap, \{\boldsymbol{W}\} \rangle$ *is a commutative monoid;*
3. $A \sqcap (B \cup C) = (A \sqcap B) \cup (A \sqcap C)$*;*
4. $\varnothing \sqcap A = A \sqcap \varnothing = \varnothing$.

2.3 Two orders: entailment and compliance

There are two orders on the set of proposals, the semilattice order ($\geq_{\cup}$) and the divisibility order ($\leq_{\sqcap}$). If we associate entailment with the semilattice order, then entailment will allow for $\vee$-introduction, but not for $\wedge$-elimination. If we associate entailment with the divisibility order, entailment will allow for $\wedge$-elimination, but not for $\vee$-introduction. The choice is guided conceptually, by seeing what one may generally conclude from a proposal (# indicates that the entailment does not go through):

Observation 3 (Behaviour of entailment)

1. $\dfrac{\textit{Let's have coffee and a biscuit}}{\textit{Let's have coffee}}$
2. $\dfrac{\textit{Let's have coffee}}{\textit{Let's have coffee or beer}}$ #

These observations show that entailment on proposals should not allow for $\vee$-introduction, but for $\wedge$-elimination.

Hence, we associate entailment with the divisibility order, i.e., the order with respect to which $\sqcap$ is almost-but-not-quite a meet operation:

Definition 11 (Entailment)
For any $A, B \in \wp\wp \boldsymbol{W}$, A entails *B, $A \vDash B$, iff $A \leq_{\sqcap} B$ (iff $\exists C. B \sqcap C = A$).*

Note that, because pointwise intersection is not idempotent, $A \vDash B$ does not mean that after expressing A, expressing B is redundant.

The semilattice order can be interpreted as follows. If $A \geq_{\cup} B$, i.e., $B \subseteq A$, then all updates proposed by B are already proposed by A. If this is the case, we say that B *complies* with A, or that

A makes B compliant. For clarity, we associate a new symbol with the semilattice order thusly interpreted:

Definition 12 (Compliance) *A makes B compliant, $A \propto B$, iff $A \geq_{\cup} B$ (iff $B \subseteq A$).*

Compliant responses to an initiative will play an important role in our pragmatic account of the implicatures of (1) and (3) in section 3. In particular, both implicature-yielding and implicature-providing responses can be characterised by means of the notion of compliance.

We wish to emphasize that from an algebraic viewpoint, entailment and compliance are both equally fundamental notions.

2.4 Implication

For implication, it is much less clear to which expression in our metalanguage the semantics of implication should correspond. Much more than in the case of conjunction and disjunction, we believe this is a matter of technical convenience and empirical adequacy. In the present paper, we make only a semi-motivated choice and spell out some formal properties.

In InqB, although not presented here, implication requires that for every possible update with the antecedent, an update with the consequent is chosen, and that the common ground is updated in a way that effectively implements this mapping from antecedent possibilities to consequent possibilities. Following the same strategy in the unrestricted case leads to the following definition (from Ciardelli et al., 2009):

Definition 13 (Conditional proposal)
$A \Rightarrow B := \{\{w \in \mathbf{W} : \textit{for all } \alpha \in A, \textit{if } w \in \alpha \textit{ then } w \in f(\alpha)\} : f : A \to B\}$

This notion of implication has some properties that one would expect of implication. For instance, $A \Rightarrow B$ gives us a proposal C such that $A \sqcap C \vDash B$, i.e., modus ponens is a sound derivation rule. Nevertheless, unlike in classical semantics, $A \Rightarrow B$ does not in general give us the *entailment-weakest* proposal C such that $A \sqcap C \vDash B$. In fact, there is no unique such proposal. This was pointed out to me by Roelofsen (p.c.) for the original definition of entailment in (Ciardelli et al., 2009), but it holds also for the new definition of entailment adopted here:

Fact 6 (No relative pseudo-complement) *There is not generally a unique $\vDash$-weakest proposal C such that $A \sqcap C \vDash B$.*

To see this, consider a model with three worlds a, b, c, let $A = \{\{a, b\}\}$, $B = \{\{a\}, \{b\}\}$. The proposals $\{\{a, c\}, \{b, c\}\}$, $\{\{a\}, \{b, c\}\}$ and $\{\{a, c\}, \{b\}\}$ are all entailment-weakest proposals C such that $A \sqcap C \vDash B$.

We do have the following result:

Fact 7 (Singleton consequent) *If B is a singleton proposal, then $A \Rightarrow B$ is the unique $\vDash$-weakest proposal C such that $A \sqcap C \vDash B$.*

Proof sketch for reasons of space: if B is a singleton set, there is only one possible mapping from A-updates to B-updates, and we can rewrite $A \Rightarrow B = \{\{w \in \mathbf{W} : \text{if } w \in \bigcup A, \text{ then } w \in \bigcup B\}\} = \{\overline{\bigcup A} \cup \bigcup B\}$. This is just classical material implication with an extra set of curly brackets.

2.5 Semantics

To obtain InqU, we associate the basic operations of our semiring of proposals with the logical connectives.

Definition 14 (Unrestricted inquisitive semantics) *For p a proposition letter, φ and ψ formulae:*

1. $[\![p]\!] = \{\{w : w(p) = 1\}\}$;
2. $[\![\bot]\!] = \{\varnothing\}$;
3. $[\![\varphi \vee \psi]\!] = [\![\varphi]\!] \cup [\![\psi]\!]$;
4. $[\![\varphi \wedge \psi]\!] = [\![\varphi]\!] \sqcap [\![\psi]\!]$;
5. $[\![\varphi \to \psi]\!] = [\![\varphi]\!] \Rightarrow [\![\psi]\!]$.

Due to fact 7, we know that negation $\neg\varphi$, defined as abbreviating $\varphi \to \bot$, denotes the pseudo-complement of $[\![\varphi]\!]$.

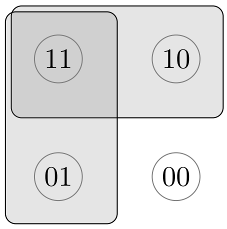
The semantics is equivalent to InqU in (Ciardelli et al., 2009), apart from the notion of entailment (and apart from some technical differences in how empty sets are treated). Classical semantics can be obtained from it by taking as semantic objects the unions of proposals:

Fact 8 (Conservativeness) *For any formula φ, $\bigcup [\![\varphi]\!]$ gives its classical meaning.*

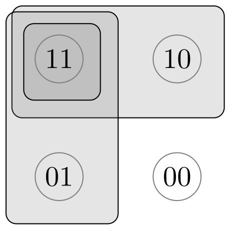
The classical meaning $\bigcup A$ of a proposal A captures the *information* provided by the proposal.

2.6 Example

As the reader can verify, (1) and (3), translated as $[\![p \vee q]\!]$ and $[\![p \vee q \vee (p \wedge q)]\!]$, are assigned the proposals depicted in figure 2.6. It is illustrated for a model that consists of four worlds (small circles), that differ with respect to two proposition letters p and q ('10' indicates that p is true and q is false, '11' that both are true, etc.). All rounded rectangles represent updates-as-states.



(a) $[\![p \vee q]\!]$



(b) $[\![p \vee q \vee (p \wedge q)]\!]$

Figure 1: The proposals denoted by (1) and (3) are distinct.

3 Unrestricted inquisitive pragmatics

Before presenting our account in detail, we will roughly sketch the *division of labour* employed to overcome the difficulties for a Gricean pragmatics mentioned in the introduction. The difficulties, recall, had to do with characterising relevant alternatives (the requirement for scales, the unwanted negation problem), the ad-hoc nature of the epistemic step, and a more general account of exhaustivity (example (3), mention-some questions).

Following (Groenendijk & Roelofsen, 2009), and in line with the collaborative nature of proposals, we assume that implicatures arise not from sentences in isolation, but from responses to an initiative. The initiator chooses which reponses are compliant, thereby *suggesting* a particular range of implicatures. The relevant alternatives for computing implicatures of a response are the update functions proposed by the initiator, in accordance with the following inquisitive version of the Gricean maxim of relation:[4]

[4]'Alternative' here is a slightly misleading notion, for some updates may be included (qua state) in others, and hence not be genuine alternatives.

Definition 15 (Maxim of relation (inquisitive))
Include in your proposals only functions that update the common ground with relevant information.

The relevant alternatives can thus be taken straight from the semantics, which renders the lexically specified, syntactic-semantic scales unnecessary, and the unwanted negation problem will not even occur.

The epistemic step will be made unnecessary by spelling out the context change potential of proposals in terms of *drawing attention* to, or away from, certain possibilities. As we will show, the pragmatic behaviour of both examples (1) and (3), as well as mention-some questions, are all captured by the same, simple reasoning pattern. In addition, a semantic characterisation of implicature-avoiding responses will be given.

3.1 Attending and unattending possibilities

We have defined InqU as a static semantics. Proposals are sets, static objects, instead of update functions on a context (although of course the elements in the set are conceived of as update functions on the common ground). Indeed, we have said nothing yet as to what kind of context proposals should change. Presumably, the context will record which proposals have been made, who committed to which update functions, etc..

For the present purposes, we assume rather minimalistically that the context change potential of a proposal is, for each proposed function, to draw attention to the possibility that it is truthfully executed, i.e., to the possibility that the actual world is contained in the update-as-state, thereby 'unattending' all previously attended possibilities. This attentive aspect of proposals occurs also, quite centrally, in (Ciardelli et al., 2009).

For responses to an initiative, we can characterise its attending/unattending potential as follows (it is more natural here to call updates-as-states *possibilities*):

Definition 16 (Attending/Unattending)
A proposal B attends *the possibilities in B. In response to a proposal A, B* unattends *the possibilities $\alpha \in A$ s.t. $\alpha \cap \bigcup B \notin B$.*

For instance, replying to $p \vee q$ with p unattends the possibility that q. Replying with the stronger $p \wedge q$ does not unattend any possibility, since *given the* *new information*, the possibility that p (q) holds is still attended.

3.2 An account of (1) and (3)

Making explicit the attentive effect of proposals reveals that answering compliantly can be partly a destructive act. Given that all possibilities raised by the initiator are relevant in accordance with the maxim of relation, unattending any one of them will require a good reason. A reasonable explanation is that the responder knows that the possibility is not, in fact, a live possibility. This reasoning pattern is spelled out below for example (1), translated as $p \vee q$, with the response p.

Reasoning pattern 2 (Unrestr. inq. account)

1. *The initiator said $p \vee q$, attending the possibilities that p and q.*
2. *The possiblities that p and q are relevant.*
3. *The responder said p, unattending the possibility that q.*
4. *The reason for unattending this relevant possibility may reasonably be that the responder believes that q is false.*

The same reasoning works for (3), translated as $p \vee q \vee (p \wedge q)$, with the response p. Now $p \wedge q$ is added among the unattended possibilities, but this makes no difference, since $\neg q$ already entailed $\neg(p \wedge q)$.

Responding to either example (1) or (3) with 'both', $p \wedge q$, does not unattend any possibility. It provides so much information that the possibilities for p and q coincide, but they are still attended given the new information (cf. definition 16). Hence, this response does not yield an implicature for either example.

Responding to the examples with $p \vee q$, however, does make a difference. In response to (1) it does not unattend any possibility and no implicature arises, whereas in response to (3) it unattends the possibility that $p \wedge q$, yielding a 'not both' implicature. We think this is as it should be.

Finally, we would like to be able to say that (1), although it does not imply anything, *suggests* that not both p and q obtain, while (3) does not. The following definition of suggestion achieves this.

Definition 17 (Suggestion)
For a proposal A and state α, A suggests *that the actual world is in α iff all singleton responses B s.t. $A \propto B$, imply that the actual world is in α.*

This definition says that any proposal suggests what its singleton compliant responses imply. Suggestions only turn into implicatures when a singleton compliant response is given.

3.3 Responses that do not unattend

There is an interesting class of non-compliant responses that explicitly avoid the 'not both'-implicature:

(6) A: Will John or Mary come to the party?
B: Well, at least John. / John, and maybe Mary too.

We say that these responses are *cautious* in the sense of choosing one compliant response, but not wishing to unattend another one.

Definition 18 (Cautious response)
B is a cautious response *to A iff for some B', B'' s.t. $A \propto B', B''$, $B = B' \cup (B' \sqcap B'')$. B is* properly cautious *iff in addition, $A \not\propto B$.*

Furthermore, a cautious response is *maximally* cautious iff one of the compliant responses it is composed of, is the original proposal itself:

Definition 19 (Maximally cautious response)
B is a maximally cautious response *to A iff for some B' s.t. $A \propto B'$, $B = B' \cup (B' \sqcap A)$. B is* properly cautious *iff in addition, $A \not\propto B$.*

Maximally cautious responses do not unattend any possibility, hence they do not give rise to an exhaustivity implicature. The reader may verify that in response to $p \vee q$, $p \vee (p \wedge q)$ is a properly, maximally cautious response (cf. example (6)).

3.4 Mention-some: relevance in interaction

In reasoning pattern 2, step (iv) is clearly the defeasible one. In particular, it relies on the assumption that the possibilities that the initiator deemed relevant, remain relevant when the responder selects one of them. Of course, this assumption is not always appropriate. In particular, in response to mention-some questions, exemplified in (4) in the introduction, selecting one possibility renders all others irrelevant (Rooij & Schulz, 2006).

For instance, in response to the first example ('I will pick up the key...'), after ascertaining that the father will be home, the possibility that the mother will be home as well is no longer relevant - one person being home is sufficient for picking up the key. Therefore, step (iv) in reasoning pattern 2 does not go through, and the response does not yield the 'not both'-implicature.

More generally, because what is relevant may change during an interaction, responses to a mention-some question do not come with an exhaustivity implicature, and hence mention-some questions do not come with an exhaustivity suggestion.

3.5 Embedded implicatures

The difficulty of embedded implicatures, recall, was that in order to get the correct implicatures for disjunctions embedded under a quantifier, the relative alternatives somehow have to be computed in the embedded position (cf. example (2)). Clearly, InqU has the advantage that alternatives are a fundamental, compositionally computed part of the semantics. Indeed, no work remains to be done except to show the present account behaves well.

As we do not wish to introduce a complete first-order machinery, we will assume a finite domain $\{d_0 \dots d_n\}$ and the existence of sufficiently many propositional variables, such that we may treat a universal quantifier as a conjunction over all individuals in the domain. This simplistic account of quantification suffices for the present purposes.

Without loss of generality, let our domain consist of Mary, John and Bob, $\{m, j, b\}$. Let k_d denote the fact that individual d read *King Lear*, and similarly o_d for *Othello*. The problematic sentence in (2) then translates as:

(7) $(o_m \vee k_m) \wedge (o_j \vee k_j) \wedge (o_b \vee k_b)$

As the reader can verify by distributing the conjunctions over the disjunctions, the proposal denoted by this formula contains an update for $o_m \wedge o_j \wedge o_b$, an update for $o_m \wedge o_j \wedge k_b$, etc..

Responding compliantly by selecting any one of these possibilities unattends all the others. By reasoning pattern 2, such responses yield the implicature that every student read only either *Othello* or *King Lear*, not both (and similar, weaker implicatures arise for non-singleton compliant responses). The formula as a whole, then, suggests exhaustivity in exactly the same way as examples (1) or (3); there is no problem with embedded implicatures.

4 Comparison to basic inq. pragmatics

We will briefly compare our approach to the inquisitive pragmatics based on InqB, developed in (Groenendijk & Roelofsen, 2009), at least as far as examples (1) and (3) are concerned. Skipping over some important, but for the present scope inessential, differences, their account of (1) could read as follows:

Reasoning pattern 3 (Basic inq. account)

1. *The initiator said* $p \vee q$.
2. p *and* q *are compliant responses, while* $p \wedge q$ *is not.*
3. $p \wedge q$ *is stronger than either* p *or* q.
4. *The reason for not making the stronger response* $p \wedge q$ *compliant might be that the initiator believes* $p \wedge q$ *to be false.*

First, note that this account, like ours, has no dubious epistemic step. Deciding to not make a stronger response compliant, like unattending a possibility in our approach, is an active deed that needs justification. Second, this account requires the assumption that relevance is closed under conjunction (for where does $p \wedge q$, as an alternative, come from?). In our account, on the other hand, what is relevant is determined solely by the initiator.

More concretely, this account fails for (or was not designed for) example (3) ($p \vee q \vee (p \wedge q)$). First, in InqB $p \vee q$ and $p \vee q \vee (p \wedge q)$ denote the same proposition. Second, transfering reasoning pattern 3 to the richer InqU would not work. For $p \vee q \vee (p \wedge q)$, step (ii) would no longer apply, and no implicature would result.

5 Conclusion and outlook

Starting from the view of meanings as proposals, we conceptually motivated and algebraically characterised an unrestricted inquisitive semantics (InqU). The algebraic backbone of InqU turned out to be a commutative, idempotent semiring, and this gave rise to a new entailment order, and a compliance order of algebraically equal stature. We hope that the algebraic characterisation of InqU will help to link inquisitive semantics to other formalisms, such as propositional dynamic logic (see Eijk & Stokhof, 2006 for a recent overview). This could lead to a transfer of many interesting results, proofs, and concepts.

Based on InqU, we defined an essentially Gricean account of some exhaustivity implicatures, and showed how it overcomes a number of difficulties for the more traditional Gricean account. Among the difficulties we discussed were the problem of characterising relevant alternatives, the problem of embedded implicatures and the unwanted negations problem. In addition, an analysis was given of the pragmatics of mention-some questions. The core ingredients for dealing with these analyses are the inherent, semantic notion of alternative in InqU and the pragmatics' focus on initiative/response pairs rather than single utterances. Both essentially followed from the same conceptual starting point: to conceive of meanings as proposals.

The present paper could not do sufficient justice to existing semantic and pragmatic theories of the phenomena discussed, several of which have already been mentioned. (Rooij & Schulz, 2006) covers most of the phenomena discussed, and a comparison between our approach and theirs will be interesting also from a methodological viewpoint: they, too, build a Gricean pragmatic theory on a non-classical semantics, in their case a dynamic semantics. The interactive view on pragmatics seems very much in line with Spector's account of exhaustivity implicatures in terms of a question under discussion (Spector, 2007). (Chierchia et al., 2008) contains many more interesting challenges for a traditional Gricean pragmatics, each of which must be investigated from the viewpoint of unrestricted inquisitive semantics and pragmatics. And there is much, much more.

For now, the relative ease (fingers crossed) with which the same reasoning scheme could be applied to the various phenomena discussed in this paper is at least a promising start.

References

Benthem, J. van. (1989). Semantic parallels in natural language and computation. In H. D. Ebbinghaus, J. Fernandez-Prida, M. Garrido, & D. Lascar (Eds.), *Logic Colloquium, Granada, 1987* (p. 31375). Elsevier, Amsterdam.

Chierchia, G., Fox, D., & Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *Handbook of semantics.* Mouton de Gruyter.

Ciardelli, I. (2010). A first-order inquisitive semantics. In M. Aloni, H. Bastiaanse, T. de Jager, & K. Schulz (Eds.), *Logic, language, and meaning: Selected papers from the seventeenth Amsterdam Colloquium.* Springer.

Ciardelli, I., Groenendijk, J., & Roelofsen, F. (2009). Attention! *Might* in inquisitive semantics. In S. Ito & E. Cormany (Eds.), *Proceedings of semantics and linguistic theory (SALT XIX).*

Eijk, J. van, & Stokhof, M. (2006). The gamut of dynamic logics. In D. Gabbay & J. Woods (Eds.), *The handbook of the history of logic. volume 6: Logic and modalities in the twentieth century* (p. 499-600). Elsevier.

Groenendijk, J., & Roelofsen, F. (2009). Inquisitive semantics and pragmatics. In J. M. Larrazabal & L. Zubeldia (Eds.), *Meaning, content, and argument: Proceedings of the ILCLI international workshop on semantics, pragmatics, and rhetoric.*

Horn, L. (1972). *On the semantic properties of logical operators in english.* UCLA.

Roelofsen, F. (2011). Algebraic foundations for inquisitive semantics. In H. van Ditmarsch, J. Lang, & J. Shier (Eds.), *Proceedings of the third international conference on logic, rationality, and interaction* (pp. 233–243). Springer-Verlag.

Rooij, R. van, & Schulz, K. (2006). Pragmatic meaning and non-monotonic reasoning: the case of exhaustive interpretation. *Linguistics and Philosophy*, *29*, 205–250.

Sauerland, U. (2005). The epistemic step. *Experimental Pragmatics.*

Spector, B. (2007). Scalar implicatures: Exhaustivity and gricean reasoning. In M. Aloni, A. Butler, & P. Dekker (Eds.), *Questions in dynamic semantics* (pp. 225–250). Elsevier.

We Did What We Could: An Experimental Study of Actuality Inferences in Dialogues with Modal Verbs

Lori A. Moon
University of Illinois at Urbana-Champaign
4080 Foreign Languages Building, 707 South Mathews Avenue
Urbana, IL 61801 USA
aralluna@gmail.com

Abstract

Modal verbs sometimes license actuality inferences, as in the sentence, *'Mayra knew that Jeff could see the mountains'*, which licenses the inference that Jeff did see the mountains. While advances have been made into modeling the interpretation of modal verbs with actuality inferences, far less has been done to determine, in a given discourse, which reading interlocutors are most likely to share. Previous work with corpora (Moon, 2011) indicated that, in uses of modal verbs in complement clauses, the type and tense of the matrix verb and the presence of intensifying adverbs (e.g., *so carelessly*) were sentence-level linguistic features which correlated with actuality inferences.This study runs and analyzes experiments with human subjects for statistically significant correlations among the features observed in corpora and the presence of actuality inferences. It is found that sentence level features do significantly bias subjects towards a particular reading and, furthermore, that the interaction between modal verbs and the types of matrix verb with which they occur is statistically significant.

1 Introduction

Theoretical semantic interpretations of modal verbs have posited multiple readings, one of which is selected by an agent based on the context in which the modal occurs (e.g., (Kratzer, 1981),(Veltman, 2005)) or the syntactic constituent to which the modal applies (e.g., (van der Auwera & Plungian, 1998),(Boland, 2006)). Although the choice among readings can be constrained by lexical properties of the modal verb, the readings available for a given modal verb still differ significantly in their temporal and inferential properties. For example, on one reading of sentence (1) below, Jeff did see the mountains,[1] and a realized ability is expressed.

1. *Mayra said that Jeff could see the mountains.*

On another reading, Mayra is being reported to have suggested that Jeff see the mountains and there is no felicity requirement that Jeff has seen the mountains before the time the utterance is made. Although sentence (1) is a completely natural sounding utterance, without additional context, it allows variability in temporal and inferential meaning which needs to be resolved in order for the correct semantic interpretation to proceed. Existing semantic theories can interpret each reading, but they have very limited ability to constrain or predict which reading interlocutors are most likely using in a given discourse.

In a discourse context which allows multiple readings of a modal verb, do sentence-level linguistic features facilitate interlocutors in conveying and interpreting matching modal verb readings? If so, does each relevant feature make a uniform contribution in biasing interlocutors towards a particular interpretation, or are there, rather, interactions among features making combinations of features have a greater effect than the sum of their individual contributions? Three specific research questions within these big questions are investigated in this study: In complement clauses such as in sentence (1) above, 1) Does the presence of a modal verb interact with the particular type of matrix (embedding) verb? 2) Does changing the tense of the modal verb from past

[1] The high likelihood of actuality inferences given sensory verbs has been observed by (Vendler, 1957) and others.

to present lower the likelihood of readings with actuality inferences across matrix verb type? and 3) Does information in the embedded clause making the actuality inference stronger (i.e., intensifying adverbs) increase the likelihood of a reading with actuality inferences equally across matrix verb type?

1.1 Background

Actuality inferences are important for natural language processing because they attribute non-modal meaning to modal verb uses.[2] For human interlocutors, an actuality inference updates the common ground of the discourse with (defeasible) information about the actual world, as opposed to the hypothetical or purely epistemic information introduced by other readings of modal verbs. Actuality inferences also carry temporal inferences with them: If an agent attributes an actuality inference to a modal verb use, it follows that the time of the described event precedes the utterance time.[3]

The problem of interpreting modal verbs in automated tasks has proven to be difficult. State of the art systems such as MacCartney 2006 and MacCartney et al. (2009:57) use lexical association with one inference pattern. The FraCas test set (Cooper et al., 1996), on which more linguistically rich entailment models are tested, only contains two uses of *would* and none with *could*. When textual inference tasks are discussed, modal interpretation is often either being overlooked or making inadequate predictions.

Uses of modal verbs with actuality inferences are particularly useful for automated inferencing systems to detect due to the information they contribute to the actual world of the discourse as well as the significantly lessened computational modeling required. The independent variables tested with subjects in this study were chosen with consideration for their tractability as encodable features for automated tasks.

[2](Hacquard, 2006), unlike (Bhatt, 1999), claims that modal verbs with actuality entailments in French and Italian do need to be modeled with possible worlds. Although her reasons could be argued to apply to the English data, they are somewhat dependent on her particular theory of event and modal scope interaction.

[3]Some accounts such as (Marneffe, Manning and Potts, 2012) treat actuality inferences as including future uses, provided there is speaker certainty (e.g., paraphrasability with *will*.)

1.2 Preliminary Work

In a previous study (Moon, 2011), 375 instances of *could* and *would* from an 80,000 line corpus of weblogs were hand-classified according to whether or not their most prominent reading was one with an actuality inference or one without an actuality inference. Within the actual or non-actual uses, examples were classified according to coarse-grained taxonomic categories such as *counterfactual, hypothetical, dispositional*, and *habitual in the past*, following various modal taxonomies (e.g., (Portner, 2009), (Ziegeler, 2000), (Celce-Murcia & Larsen-Freeman, 1999)).

Examples were sorted into syntactic categories in order to allow the classification problem to be potentially described by construction-based rules (cf Fernandez et al. 2007). The most prominent syntactic categories which emerged were: **Conditional (Cond):** instances in the consequent clause of a conditional sentence. **Free-Standing (FS):** instances in a main clause, possibly with adjuncts or coordination. **Relative Clause (RC):** instances in a clause headed by a noun phrase. **Complement Clause (CC):** instances in a clause which is the complement of a verb. **Other (OT):** instances in adjunct or coordinating constructions which seemed to exhibit exceptional properties.

For sentences with the modal verb *would*, the syntactic construction was found to be a strong factor in determining whether a given use of the modal was likely to have an actuality inference or not. For *could*, however, the syntactic construction was less helpful. As shown in figure 1, many constructions were almost equally likely to encode actual or non-actual uses.

Additional feature exploration was done in the corpus data within construction type in order to determine which linguistic expressions correlated with actuality inferences. For complement clauses with *could*, certain properties of the matrix verb type, matrix verb tense, and intensifying adverb phrases appeared to correlate with actuality entailments.

In this study, the generalizations from corpus data mentioned in (Moon, 2011) are treated as independent variables in an experimental study with human subjects. It is proposed that, if the same features which emerged in corpora are varied in controlled linguistic contexts with statistically significant correlations to the presence of ac-

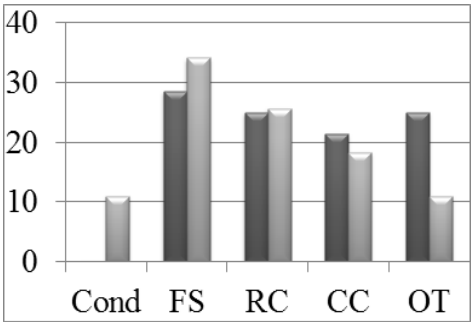


Figure 1: Given an actual occurrence of the modal *could* (black) or a non-actual occurrence (gray), this chart shows the likelihood that a particular construction will be used (figure from Moon 2011). Cond = Conditional, FS = Main Clause, RC = Relative Clause, CC = Complement Clause, OT = Other.

tuality inferences, their validity as features determining the modal verb reading is strongly corroborated.

2 Research Question 1: Matrix Verb Type and Modal Verbs

It was observed in corpus data (Moon, 2011) that the matrix verb type appeared to be a strong determining factor in actuality entailments. Although Levin verb classes (Levin, 1993) were first considered, it was the three general classes defined in Karttunen's work (Karttunen, 1973), (Karttunen, 1974) which emerged as forming separate categories with respect to actuality inferences.

Karttunen proposed that three types of matrix verbs respond differently regarding whether or not they allow presuppositions in their scope to project. Factive verbs, and most other verbs, are considered to be *holes* to presuppositions, allowing presuppositions in their scope to percolate. Verbs of saying and telling are considered to be *plugs*, relativizing any presuppositions in their scope to the beliefs of the reported speaker. Propositional attitude verbs, such as *believe*, are claimed to alternatively allow or disallow the percolation of presuppositions in the clauses in their scope. Karttunen's work on presupposition projection remains very influential in both implementations (e.g.,(Clausen & Manning, 2009)) and experimental work (e.g., (Smith & Hall, 2011)).

Actuality inferences would be difficult to categorize as presuppositions. In English, they are pervasively cancelable: Allowing continuations which re-construe the modal reading as one without an actuality inference as in sentence (2):

2. *Mayra knew that Jeff could see the crime taking place, {but he was actually sleeping at the time/if he wanted to}.*

To speak of actuality inferences projecting through embedded clauses would also not be accurate, since the unembedded modal clause in itself rarely has an actuality inference. Therefore, Karttunen's theory is not straightforwardly tested in this study, rather, it provides a classification of embedding verb types which has proven helpful for presupposition projection and also correlates with presence or absence of actuality inferences.[4]

By alternating the matrix verb class with the presence or absence of a modal verb, we test whether the verb classes behave as holes, plugs, or filters regarding the assertion in their scope in cases such as sentence (3) where the modal is omitted.

3. *Mayra {knew/thought/said} that Jeff saw the crime taking place.*

When the modal is included, as in sentence (2) above, there are two possible outcomes: The first hypothesis is that the modal verb will make a uniform contribution across matrix verb type in lowering the likelihood of a reading with actuality inferences. In this case, the contribution of each feature is uniform and can be composed to represent the overall likelihood of a particular reading. An alternative hypothesis is that the presence of a modal verb will have a greater effect with some matrix verb types than with others. If the second hypothesis holds, then it follows that the contribution of the modal verb and the contribution of matrix verb type in determining a modal verb reading cannot be described by any simple combination of the contribution of each part. Rather, the semantic theory must represent the fact that the interaction of modality and matrix verb type is greater than the combination of its parts.

[4] One application of this observation is that, in automated systems such as that of (Clausen & Manning, 2009), the regular expressions used to locate factive verbs for representing presupposition projection can be used to detect embedding verbs with modal verbs in their complement clause and encode predictions regarding actuality inferences.

3 Research Question 2: Past under Past Embeddings versus Past under Present Embeddings

There are various phenomena which are said to exhibit *sequence of tense* (sot) effects (Abusch, 1997), (Ogihara, 1995). This study examines one specific type of sot phenomena: Past under past embedded clauses in which the central past tense modal verb, *could* (Quirk et al., 1985) (Crouch, 1993) occurs under a past tense embedding verb.

The modal verb *could* allows a *simultaneous reading*: A reading in which the reported event occurred at the time of the subject of the main clause's reported speech act:

4. *Yvette said that Jill could jump on the trampoline for twenty minutes.*

In sentence (4), there is a simultaneous reading on which Jill has the ability to jump on the trampoline for twenty minutes, also called the *generic* reading (Bhatt 1999:173). At the time of Yvette's reported speech act, Jill had the ability to jump on the trampoline for twenty minutes. It is even possible for the (defeasible) inference to be made that Jill has, at the time of Yvette's saying that she could jump on the trampoline for twenty minutes, already jumped on the trampoline for twenty minutes in the past at least once.

Sentence (4) also involves a *sequential reading*: A reading in which the potential event of Jill jumping on the trampoline for twenty minutes occurred before Yvette's act of reporting Jill's ability. This is sometimes called an *episodic* reading (Bhatt 1999:173).

Cross-linguistically, some languages which have morphological aspectual marking on their modal verbs distinguish generic from episodic readings in that only the latter can occur with the perfect aspect (Bhatt, 1999). In English, however, both the episodic and generic readings of sentence (4) license actuality inferences, provided that, on the episodic reading, the utterance time is taken to be after the time of the potential event in the scope of the modal (i.e., Jill's jumping on the trampoline for twenty minutes). Such readings might be described as being *metaphysical* or *historical* (Condoravdi 2002:77-84)[5] in that they describe a past prediction about an event which was future at the time of reported speech, but past relative to the time of the utterance. The potential event in the scope of the modal might or might not have occurred at the speech time. On the reading in which Jill did jump on the trampoline for twenty minutes, the prediction of the reported speaker held and, on the reading in which Jill did not jump on the trampoline, the prediction of the reported speaker has proven to be incorrect. Only the former reading licenses an actuality entailment.

It is hypothesized that, when subjects accept an actuality inference in past under past embeddings, they are accepting a historical (or possibly generic) reading. In accepting a historical reading, they are rejecting the reading in which the past under past sentences describes a prediction about an event which follows the speech time.

In past under present embeddings, the three verb classes are still used, but with the embedding verb is in the present tense rather than in the past tense. As discussed in section 4.1, past under past embeddings allow various sot phenomena, each of which locates the embedded (telic) event as preceding the speech time. It is hypothesized that, by using a present tense embedding verb, such as sentence (5), the modal will be interpreted as referring to a future prediction rather than a comment on a past event.

5. *Rika understands that Jordan could report her to the authorities.*

The modal in sentence (5) most prominently describes a hypothetical future possibility rather than Jordan's past ability.

Given the readings available, it is hypothesized that readings with actuality inferences do not occur in past under present embeddings. This effect is hypothesized to be uniform across matrix verb type.

4 Research Question 3: Intensifying Adverb Phrases

One feature which emerged and which has not received a lot of attention with respect to actuality entailments were intensifying adverbs such as *so emphatically* or *so carelessly*. Such expressions were used in the corpus data most frequently with actual readings. In general, intensifying adverb constructions modify an actual event. However, non-actual readings with *so + ADVERB* are licit

[5](Condoravdi, 2002) does not include *could* among her modals for the past, however, it seems that this reading of *could* is included in her description of metaphysical uses of non-root modals (2002:78).

as in sentence (6):

6. *Judson believes that Imri could so quickly forget his parents.*

Sentence (6) can be uttered, for example, as a speculation about Imri's anticipated mental state during his future absence. The reading without an actuality inference, without a background context, does not, however, seem to be the most prominent one.

The presence of intensifying adverb phrases is hypothesized to increase the number of subjects accepting readings with actuality inferences across matrix verb type.

5 Methods

It was not feasible to test every possible value of the proposed features (3 x 2 x 2 x 2 = 24 conditions), so only select combinations were tested, resulting in twelve test conditions:

(ia) factive verb, past, no modal: Marian was amazed that Dewayne silenced the witness.
(ib) factive verb, past, modal: Marian was amazed that Dewayne could silence the witness.
(ic) factive verb, past, modal, so-phrase: Marian was amazed the Dewayne could so effectively silence the witness.
(id) factive verb, present, modal: Marian is amazed that Dewayne could silence the witness
(iia) saying verb, past, no modal: Marian said that Dewayne silenced the witness.
(iib) saying verb, past, modal: Marian said that Dewayne could silence the witness.
(iic) saying verb, past, modal, so-phrase: Marian said that Dewayne could so effectively silence the witness.
(iid) saying verb, present, modal: Marian says that Dewyane could silence the witness.
(iiia) propositional attitude verb, past, no modal: Marian believed that Dewayne silenced the witness.
(iiib) propositional attitude verb, past, modal: Marian believed that Dewayne could silence the witness.
(iiic) propositional attitude verb, past, modal, so-phrase: Marian believed that Dewayne could so effectively silence the witness.
(iiid) propositional attitude verb, present, modal: Marian believes that Dewayne could so effectively silence the witness.

5.1 Subjects

A total of 41 subjects were tested, with five results discarded due to acquisition of English as their primary language later than three years of age. Subjects were recruited through two introductory linguistics courses as well as through the author's personal contacts. All subjects were undergraduate students at the University of Illinois at Urbana-Champaign except one individual living in another US city. Twenty-six of the subjects were female and ten male. The age range was from 18-41 years old (average 21). All but one subject reported their English proficiency as level '5' on a scale from 1-5 where 1 was 'beginner' and 5 'advanced' (one reported 4). Two bi-lingual speakers considered themselves to be more proficient in a language other than English (Spanish and German). Their results were included since they claimed to use English as their primary language prior to age three and self-reported their proficiency as 'advanced'. Other languages spoken by subjects but considered less primary than English included Danish, French, Hindi, Japanese, Korean, Mandarin Chinese, Taiwanese, and others. Only four subjects were not born in an English-speaking country but moved to one by age three.

5.2 Apparatus

Experimental results were collected via online surveys through SurveyGizmo Student Edition. Subjects completed the survey in two parts which could be taken at separate times at the subjects' convenience. Thirty-six subjects were tested via online surveys (12 scripts, 3 subjects each). Subjects were presented with a scenario containing the target sentence and asked whether the predicate embedded under the scope of the modal had already occurred at the time when the dialogue was taking place. Subjects answered 'yes' or 'no' and then rated their certainty on a five-point scale.

Finding a context in which subjects can potentially read the modal as having an actuality inference and potentially read it as being predictive is difficult. Each dialogue involved a scenario in which two agents, Ann and Jacob, took discourse turns. Ann's role was to present a topic which Jacob then commented on either requesting more information or affirming what was said. The target data were always presented preceded by the discourse particle, *well*. An example is in figure (2).

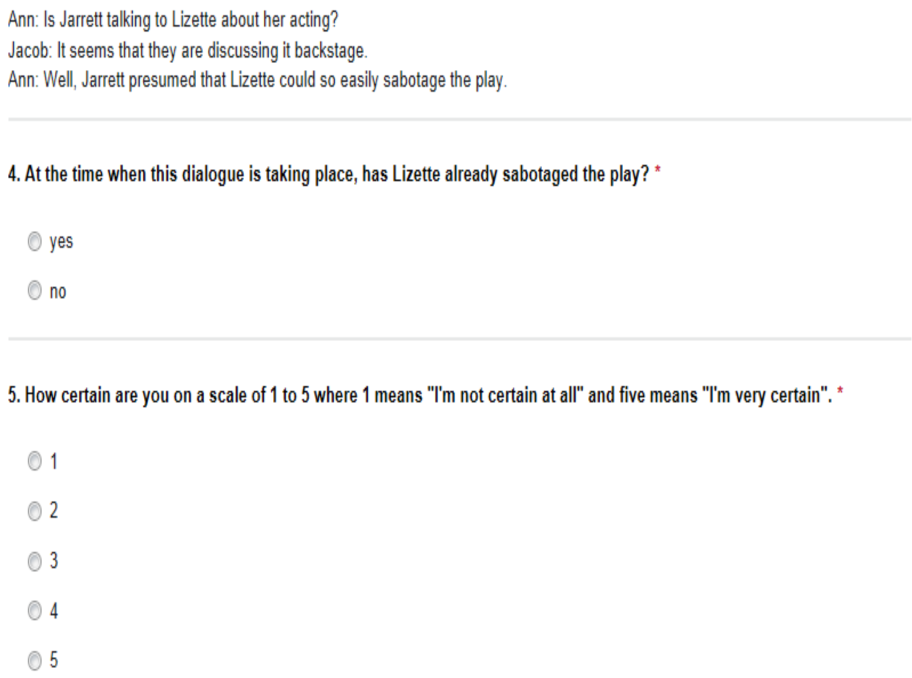


Figure 2: Example of SurveyGizmo interface

In each of the 36 scenarios (three examples of each of the twelve conditions tested) in which the independent variables were manipulated, there was, at a minimum, a historical and a predictive reading available.[6]

On the historical reading of the scenario in figure (2), Ann is reporting Jarrett's presumption after the time during which Lizette could have sabotaged the play. The discourse particle is read as introducing an affirmation that Jarrett's presumption has some relation to the present unspecified state of affairs. On a predictive reading, the discourse particle is read as providing additional information explaining the situation discussed in the first and second turns.

By the use of *already* in the question, the actuality inference is suggested to the subject. However, for it to be possible, the subject must read the modal report as historical, not predictive. If the subject can only get a predictive reading, then she will not respond that there is an actuality inference. This bias was the same across conditions.[7]

6 Results

Results were run for 'yes' or 'no' responses only (with degree of certainty measured separately), as shown in figures 3 and 4 . A sample sentence is given for each of the twelve conditions and indexed in the results table in figure 3 in section 5 above.

Figure 4 shows the basic trajectory of subjects' responses by showing the percentage of 'YES' responses to each condition.

[6] The complete set of scenarios used in testing is available at https://netfiles.uiuc.edu/lcoulter/ExperimentalScripts.

[7] When subjects in a pilot study were presented with sentences without a discourse context, particular problems were encountered with future readings. The pilot used two other conditions: past temporal adverbs and future temporal adverbs. It was predicted that future temporal adverbs, as in *Mary said that Jarrett could cook the fish tomorrow*, subjects would overwhelmingly report no actuality inference. However, it was found that a significant number of subjects accepted an actuality inference. It was surmised that subjects attributed an additional level of evidentiality to the data, assuming that whoever was reporting the data was reporting it after the quoted *tomorrow* had passed. Perhaps some notion of Grice's relevance maxim (Grice, 1975) can then explain the assumption subjects made that there would be no point in reporting an overtly future prediction unless it had come to pass. The addition of a brief dialogue as a control condition helped ensure that subjects evaluated the possibility of an actuality inference at the utterance time.

	(a) past, -modal		(b) past, +modal		(c) past, +modal, so		(d) present, +modal	
	YES	NO	YES	NO	YES	NO	YES	NO
(i) Factive	**100%**	**0%**	**61.11%**	**39.81%**	**75.92%**	**24.07%**	**49.07%**	**50.92%**
certainty (stdev)	4.69 (0.55)	N/A	3.93 (1.09)	3.69 (0.61)	4.22 (0.79)	4.18 (0.46)	4.13 (0.81)	4.18 (0.46)
(ii) Say/Tell	**97.22%**	**2.77%**	**18.51%**	**81.48%**	**47.22%**	**52.77%**	**13.88%**	**86.11%**
certainty (stdev)	4.51 (0.75)	1.33(n/a)	3.60 (1.38)	3.73 (1.05)	3.71 (0.96)	3.42 (1.32)	3.88 (0.8)	3.75 (0.92)
(iii) Propositonal	**54.62%**	**45.37%**	**18.51%**	**81.48%**	**31.48%**	**68.51%**	**2.77%**	**97.22%**
certainty (stdev)	3.97 (1.06)	3.4 (1.04)	3.58 (0.87)	3.85 (1.1)	3.69 (1.02)	3.57 (1.16)	3 (2.82)	4.20 (0.90)

Figure 3: Results of study of complement clauses with *could*. Alphanumeric labels correspond to sample conditions

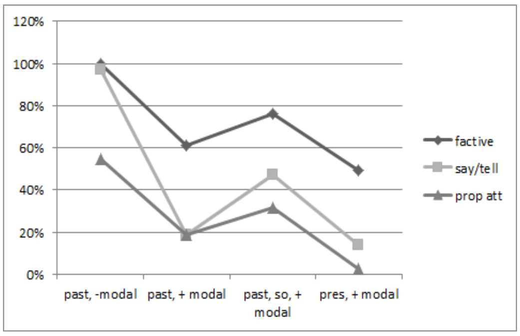


Figure 4: Plot of percentage of subjects accepting actuality inferences by condition and verb type. Y axis = percentage of subjects giving a 'yes' response (e.g., 100% means all subjects responded with 'yes' to all three samples of the condition).

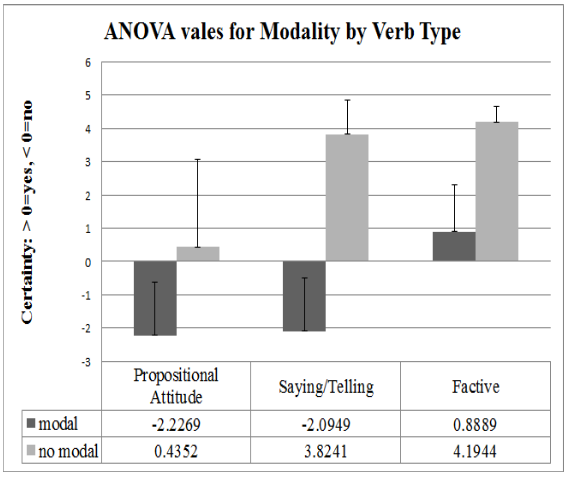


Figure 5: Effects of the modal verb *could* on actuality inferences by verb class.

6.1 Matrix Verb Type and Modal Verbs

Regarding the research question of how the matrix verb interacts with the presence or absence of a modal verb, it was found that assertions project through verbs of saying or telling on par with holes such as factives.

Column (a) in figure 3 is taken to illustrate the isolated effects of past under past embeddings in which no modal is present. Sot are known to occur in such contexts, which allow a sequential or a simultaneous past reading. Without the modal present, predictive readings would not be possible. The internal argument of the embedded verb referred to a telic event in all the data, so habitual readings were less likely.

The results are that factive verbs (YES: 100%) and verbs of saying and telling (YES: 97.2%) result in subjects almost always reporting actuality inferences, but propositional attitude verbs show a near even split among interpretations (YES: 54.62%). These results are somewhat surprising compared to those predicted in (Karttunen, 1973), (Karttunen, 1974) for presupposition projection. Without the presence of a modal verb, an actuality entailment exists in the embedded clause, however, subjects do not treat information in the scope of verbs of saying and telling as less actual. Whether human subject judgments on presuppositions are the same in this context is a topic for future experiments. As will be seen in the results of column (b), introducing uncertainty in the embedded clause drastically changes subjects' responses.

For column (b), past under past embeddings with *could*, the percentage of subjects making an actuality inference greatly decreases for all embedding verb types, but much more sharply for verbs of saying and telling (NO: 81.48%), which pattern identically to propositional attitude verbs (NO: 81.48%). Factive verbs remain majority actual (YES: 61.11%).

Column (b) gives evidence for the role of evidentiality in interpretation: When another person's views are words are being reported, the presence of the modal makes the report epistemically uncertain. On epistemically uncertainty readings of *could*, it is the subject of the main clause who is unsure, but it can often be inferred that the person reporting the speaker's view is epistemically uncertain. The effects of evidentiality can be seen if co-indexing occurs as in *'I said that I could cook the fish'* which can be easily read with the actuality inference that the speaker actually did as she said.

For ANOVA analysis, the binary 'yes' or 'no' responses and five-point certainty scale were converted to an even ten-point continuum of 0.5-interval values from -4.5 to 4.5 ('very certain' 'no' and 'very certain' 'yes', respectively).[8],[9]

A 2x3 ANOVA was run crossing modality (2 levels: Presence or absence of the modal) and verb class (3 levels: Factive, saying or telling, and propositional attitude), as show in categories (i-iii) a-b of figure 3.

This measure was run to determine the effect of the presence of the modal on the interpretation of the three verb classes as shown in figure 5.

The verb type was found to be statistically significant $(F(2, 70) = 119.6; p < 0.001)$ with factive verbs raising the likelihood of assertions projecting and propositional attitude verbs lowering the likelihood. The presence of a modal verb was also statistically significant $(F(1, 35) = 319.07; p < 0.001)$ in lowering the likelihood of actuality inferences across matrix verb type. The interaction between verb type and the presence of the modal verb was also statistically significant $(F(2, 70) = 53.55; p < 0.001)$, with verbs of saying or telling being affected most strongly by the presence of a modal verb greatly lowering their likelihood of being interpreted as having an actuality inference.

6.2 Past versus Present Tense Matrix Verb

Column (d) in figure 3 sought to see to what extent the tense of the embedding verb interacted with actuality inferences in complement clauses. It was found that, under present embeddings, about half of the subjects still accepted the actuality inference under factive verbs (NO: 50.9%). However, with verbs of saying or telling (NO: 86.11%) and propositional attitude verbs (NO: 97.22%) the percentage of subjects rejecting actuality inferences increased to nearly absolute agreement.

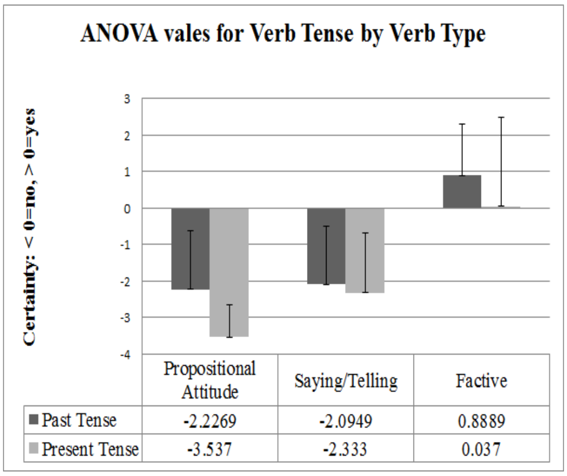


Figure 6: Interaction of tense with other features.

Column (d), along with column (b), show that the combination of verb tense and embedding verb type are the strongest features determining whether or not a modal verb licenses a prominent reading in which an actuality inference holds.

A 2x3 ANOVA was run crossing tense of the matrix verb (2 levels: Past and present) and matrix verb type (3 levels: Factive, saying or telling, and propositional attitude) as in (i-iii), (b and d) of figure 3. The measure did not show a statistically significant interaction between verb type and tense (6), but both tense and verb type were found to be statistically significant in themselves $(F(1, 35) = 11.68; p < 0.003, F(2, 70) = 76.08; p < 0.001$, respectively) in that present tense lowers the likelihood of a reading with an actuality inference.

6.3 Intensifying Adverbs

Column (c) in figure 3 adds an intensifying adverb phrase to the past under past modal embedding to see if sentence-level features in the embedded clause can coerce subjects to make actuality inferences. If so, then it shows more clearly which embedding verbs strongly prevent actuality inferences.

It was found that all embedding verb types led more subjects to accept actuality inferences, but only factives (YES: 75.92%) were above the 50% threshold. Verbs of saying were very close to an

[8] The resulting ten point scale placed the origin at the mid-point between an answer of 'yes' with certainty of one and an answer of 'no' with certainty of one.

[9] See (McKinstry, Dale, and Spivey, 2008) for experimental evidence that the divide between 'yes' and 'no' responses is more of a continuum than binary.

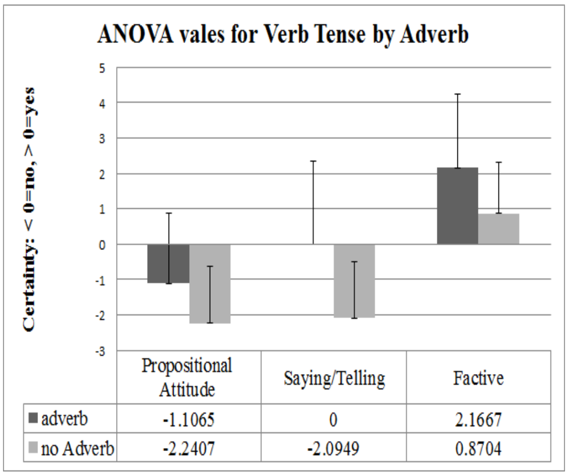


Figure 7: Interaction of 'so + adverb' phrases with other features.

even split (NO: 52.7%), with propositional attitude verbs remaining least actual (NO: 68.51%).

A 2x3 ANOVA of intensifying adverb phrases (2 levels: Presence or absence of the adverb) given verb type (3 levels: Factive, saying or telling, and propositional attitude) was also run. This measure showed the adverb to be statistically significant ($F(1, 35) = 31.05; p < 0.001$) in raising the likelihood of a reading with an actuality inference when it was present. Verb type ($F(2, 70) = 94.68; p < 0.001$), again, was statistically significant, but the interaction of the two was not statistically significant.

7 Discussion

Although English does not have overt morphological markings helping distinguish uses of modal verbs with actuality inferences from uses without, there are a number of sentence level linguistic features which assist in limiting the most likely reading. Furthermore, when embedded in a discourse context allowing both historical and predictive readings, subjects systematically interpret the discourse markers and context in a way that facilitates the most likely reading of the modal based on sentence level features.

Matrix verb type was the most prominent feature in determining actuality inferences. This fact cannot be straightforwardly explained by the theory from which the verb classes were chosen (Karttunen, 1973), (Karttunen, 1974), but it rather shows that levels of evidentiality and intuitions about the limitations of epistemic knowledge influence modal verb readings.

The second most prominent determining feature was the past tense of the main verb. Sot phenomena are predicted to occur in past under past embeddings, making the possibility of actuality inferences clear. However, the uniform trajectory across verb type, as shown in figure 4, indicates that embedding verbs play a much stronger role than tense in determining the reading.

In response to the big research questions, it was found that sentence level features do indeed help determine which reading interlocutors are most likely to attribute to a given use of a modal in that the features tested: Matrix verb type, matrix verb tense, and intensifying adverbs all had a statistically significant effect on the modal verb reading.

It was found that some features have a uniform contribution to what reading interlocutors are most likely to choose. When the matrix verb is in the present tense, all matrix verb types have a lower likelihood of being interpreted as having actuality inference. Similarly, intensifying adverbs have an effect across verb type of increasing the likelihood of a reading with actuality inference.

Other features, such as the interaction between modal verbs and the matrix verb type can not be represented as a simple combination of the matrix verb type combined with the contribution of the modal. Rather, verbs of saying or telling are much more affected by the presence of a modal than factive or propositional attitude verbs. These results present particular challenges for any theory attempting to represent a probabilistic bias towards one reading in terms of uniform contributions of constituents in a compositional semantic approach.

The results of our study have an impact on theoretical models of the role of sentence level features on modal verb interpretation in discourse. They also have applications for automated models of modal verb interpretation, providing tractable features correlating with a strong likelihood of modal verb readings with actuality inferences.

In future work, we will test features correlating with actuality inferences in other syntactic constructions and with other past central modal verbs. We will also work to improve our discourse scenarios in order to measure when human subjects make actuality inferences rather than measuring whether they accept or reject them when they are suggested.

Acknowledgments

The author's committee was helpful in supervising this research: Professors Tania Ionin, Shalom Lappin, Peter Lasersohn, and José Meseguer. Professor Michael Spivey's workshop at the 6th annual Experimental Methods in Computational Linguistics boot-camp was indispensable in learning to run ANOVA on the data. The audience at NASSLLI 2012 who attended the related talk at Reasoning and Interaction (RAIN) gave excellent feedback as well as Professor Elizabeth Allyn Smith, and members of the University of Illinois at Urbana-Champaign Experimental Pragmatics Group. Finally, warm thanks go to the anonymous reviewers for helpful suggestions and encouragement on this project.

References

Dorit Abusch. 1997. Sequence of Tense and Temporal de Re. *Linguistics and Philosophy*, 20(1): 1-50.

Rajesh Bhatt. 1999. *Covert modality in non-finite contexts*. PhD thesis, University of Pennsylvania.

Annerieke Boland. 2006. *Aspect, Tense and Modality: Theory, Typology, Acquisition, Volume I.* Center for Language and Communication.

Marianne Celce-Murcia and Diane Larsen-Freeman. 1999. *The Grammar Book: An ESL/EFL Teacher's Course.* Heinle and Heinle Publishing Company. Boston, MA.

David Clausen and Christopher D. Manning. 2009. Presupposed content and entailments in natural language inference. *Proceedings of the 2009 Workshop on Applied Texual Inference, ACL-IJCNLP 2009.* pages 70-73.

Cleo Condoravdi. 2002. Temporal Interpretation of Modals: Modals for the Present and for the Past. Chapter in *The Construction of Meaning*, pages 59-88. CSLI Publications.

Robin Cooper, Dick Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. *Using the Framework*. The FraCaS Consortium, January 1996.

Richard Crouch. 1993. The temporal properties of English conditionals and modals. PhD thesis, University of Cambridge.

Raquel Fernandex, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397-427.

H. P. Grice. 1975. Logic and conversation. *Syntax and Semantics, 3: Speech Acts*, ed. by P. Cole & J. Morgan. New York: Academic Press.

Valentine Hacquard. 2006. Aspects of modality. PhD thesis, Massachusetts Institute of Technology.

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

Lauri Karttunen. 1974. Presupposition and linguistic context. *Theoretical Linguistics*, 1:181-194.

Angelika Kratzer. 1981. The notional category of modality. Chapter in *Words, Worlds and Contexts*.

Beth Levin. 1993. *English Verb Classes and Alternations*. University of Chicago Press.

Bill MacCartney. June 2009. Natural language inference. PhD thesis, Stanford University.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid texual entailments. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 41-48.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts 2012 Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301-333.

Chris McKinstry, Rick Dale, and Michael J. Spivey 2008 Action Dynamics Reveal Parallel Competition in Decision Making. *Psychological Science*, 19.1:22-24.

Lori Moon. 2011. Discriminating among 'actual' and 'non-actual' uses of *could* and *would* for improved text interpretation. *Building Representations of Common Ground with Intelligent Agents: Technical Report from the AAAI Fall Symposium*.

Toshiyuki Ogihara. 1995. Double-access sentences and references to states. *Natural Language Semantics*, 3:177-210.

Paul Portner. 2009. *Modality*. Oxford University Press.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartivik. 1985 *A Comprehensive Grammar of the English Language*. Longman.

Elizabeth Allyn Smith and Kathleen Currie Hall. 2011 Projection Diversity: Experimental Evidence. Proceedings of the Workshop on Projective Meaning, ESSLLI 2011.

Johan van der Auwera and Vladimir Plungian. 1998 Modality's Semantic Map. *Linguistic Typology*, 2:79-124.

Frank Veltman. 2005 Making Counterfactual Assumptions. *Journal of Semantics*, 22.2:159-180.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143-160.

Debra Ziegeler. 2000. The role of quantity implicatures in the grammaticalisation of *would*. *Language Sciences*, 22:27-61.

Children learn Language in Conversation

Eve V. Clark
Department of Linguistics
Margaret Jacks Hall, Bldg 460
Stanford University
Stanford, CA 94305-2150
USA
`eclark@stanford.edu`
`http://www.stanford.edu/~eclark/`

In learning a language, children need to learn both the conventional forms for their community, and how to use those forms. I will argue that children learn language – both forms and uses in the course of conversation.

I will draw on findings from adult offers of new words and evidence for children's uptake; from adult reformulations of child errors and children's attention to these reformulations, and from adult scaffolding of children's early 'narratives', where children rely on adult knowledge of an event in order to make their contribution.

Cues to turn boundary prediction in adults and preschoolers

Marisa Casillas
Stanford University
Department of Linguistics
middyp@stanford.edu

Michael C. Frank
Stanford University
Department of Psychology
mcfrank@stanford.edu

Abstract

Conversational turns often proceed with very brief pauses between speakers. In order to maintain "no gap, no overlap" turn-taking, we must be able to anticipate when an ongoing utterance will end, tracking the current speaker for upcoming points of potential floor exchange. The precise set of cues that listeners use for turn-end boundary anticipation is not yet established. We used an eyetracking paradigm to measure adults' and children's online turn processing as they watched videos of conversations in their native language (English) and a range of other languages they did not speak. Both adults and children anticipated speaker transitions effectively. In addition, we observed evidence of turn-boundary anticipation for questions even in languages that were unknown to participants, suggesting that listeners' success in turn-end anticipation does not rely solely on lexical information.

1 Introduction

Turn-taking in human communication is efficient: we usually switch between speakers with brief pauses. Though there is a wide distribution of gap lengths in everyday conversation, the median gap between conversational turns is close to zero milliseconds, and maintaining brief inter-speaker junctions may be universal to human languages (de Ruiter et al., 2006; Heldner and Edlund, 2010; Stivers et al., 2009). These gaps, though brief, result in minimal overlap, and beg the question of how we manage to come in with such precise timing.

Sacks, Schegloff, and Jefferson (1974) noted that inter-speaker gaps are too brief for listeners to be relying on turn-end silences before starting up their response. They suggested that instead we track ongoing turns for cues to their eventual end, using linguistic information about syntactic, propositional, and intonational structure. Using these cues, listeners should be able to predict the moment at which a speaker will stop speaking with high accuracy. This insight was important, but they did not further investigate which cues—whether linguistic or non-linguistic–listeners track.

More recent research has addressed this question, investigating which linguistic cues might be most informative in anticipating the close to an ongoing turn. Corpus study of available cues has yielded somewhat inconclusive results since so many linguistic boundaries co-occur (Caspers, 2003; Ford and Thompson, 1996). Even if reliable turn-end cues were apparent, we could not be confident that listeners actually attended to them to in conversation without experimentally manipulating them and measuring their effects on listeners.

De Ruiter and colleagues (2006) created an experimental paradigm to measure turn boundary anticipation while also beginning to test which cues were most informative in this process. They extracted utterances from a recording of a spontaneous conversation and presented them to participants over headphones. Participants were asked to press a button at the moment they anticipated the speaker would stop speaking. Participants were extremely accurate in identifying the moment before a turn was about to end. To test the effects of different cues on anticipation, they separately controlled for the presence of intonation and lexical information. There was no significant difference between participants' accuracy when intonation was present and when it was omitted from the stimulus. When lexical information was taken away, however, participants' accuracy declined significantly. De Ruiter et al. thus suggested that word-level information is of primary importance in turn-boundary anticipation.

Although the de Ruiter study was carefully controlled, the button-pressing task was explicit, and might easily have focused participants' attention on words and word-level information more so than they would have been otherwise, especially since the instructions asked for precisely-timed responses. If this were the case, their results would reflect a use of linguistic cues under somewhat unnatural conditions. In addition, de Ruiter et al. (2006) did not control for all prosodic cues—duration was left unmodified in their stimuli. This information might have accounted for some of their accuracy effects in the condition without intonation.

Many people have the intuition that intonation and rhythm are part of the prediction process, but may be more important prior to the end of the turn, at which point lexical information may be most informative. Carlson, Hirschberg, and Swerts (2005) showed that listeners can use prosodic cues to predict the strength of upcoming prosodic breaks. The estimation of upcoming prosodic breaks can help listeners determine when a speaker-switch will be appropriate, even without lexical information (Carlson et al., 2005; Heldner et al., 2006). These experiments were run on "offline" judgments, unlike those in the de Ruiter et al. (2006) study—which found no prosodic effects. Could prosodic effects emerge during online speech processing under different experimental circumstances?

Our current work uses eye-tracking as an implicit measure of turn boundary anticipation. This method allows us to study both adults and children and to systematically manipulate the content of the videos we track.

Tice & Henetz (2011) explored eyetracking as a possible alternative method for measuring online turn processing, which they call Observer Gaze. They seated participants in front of a large screen, under which was tucked a small digital video camera tilted toward participants' faces. While viewing a one-minute dyadic, split-screen conversation in English, participants consistently tracked the current speaker with their gaze. In addition, they anticipated the ends of turn boundaries by looking at the next speaker on question-answer pairs. Observer Gaze is founded upon natural looking behavior—observers tend to look at the current speaker during his or her turn (Kendon, 1967; Bavelas et al., 2002). It requires little or no instruction and allows experimenters to collect high temporal resolution looking data over the course of a conversation. Thus, this method provides a measure of turn-boundary anticipation that we can use to investigate the cues that contribute to this ability.

Since it is a passive method founded on natural, spontaneous behavior, Observer Gaze can be used with both child and adult participants to begin exploring the developmental trajectory of turn-end boundary prediction. We are interested in comparing adult turn-end prediction skills with those of children because of the protracted development of turn-taking. By age five, children's turn-taking skills are still not up to the timing standards of adults. Even in adjacency pairs, when the response is often restricted and the context makes clear who the next speaker is, children's responses are still delayed. It has been proposed that their delay is due to complexity and predictability level of responding to the question at hand (Garvey and Berninger, 1981; Casillas et al., in preparation), but we do not yet know whether children's delay is due to the need to formulate a response or a slowly developing ability to predict turn-end boundaries. The eye-tracking method described above makes it possible to compare adults and children directly, allowing for investigation of this question in our study.

In the current study, we introduce a simple method for controlling word-level information in the speech signal: we show participants videos of languages that they do not speak. Though the non-lexical signals in the videos (e.g., intonation, prosody, gaze, gesture) are foreign to the participants, the information may still be robust enough to support online turn-tracking. Because the linguistic cues are foreign, eye gaze behavior while watching a foreign language (which has similar, but not identical cues) is a stringent test of the use of non-lexical cues in online-turn-processing. To keep the stimuli engaging for children, we used child-oriented speech (as described below) in the video stimuli.

2 Methods

2.1 Participants

Seventy-two pre-school aged children (19 three-year-olds, 32 four-year-olds, and 21 five-year-olds) and 11 adults participated in the study. All were native speakers of English who had little to no language experience with the four non-native languages used in the stimuli (see Procedure below).

2.2 Materials

The video segments were recorded in a sound-attenuated booth by two native speakers of each language (all non-native English speakers enrolled in graduate study in the U.S.) Each person was audio recorded from a lapel microphone (one on the right channel and one on the left) feeding into a Marantz PMD 660 solid state field recorder. Participants were video-recorded from the iSight of a MacBook. Pairs of speakers were selected by native language, and ranged from acquainted individuals to good friends. They were asked to speak on four topics for 20 minutes (five minutes each on favorite foods, entertainment, hometown layout, and pets). Following this recording they were asked to choose a topic relevant to young children (e.g., riding a bike, eating breakfast, siblings) and improvise on that topic as if they were on a children's television show until they had at least 30 seconds of continuous material. Most pairs took less than three minutes to record these "child-friendly" improvised conversations, and the resulting recordings remained natural but engaging for both young children and adults. The audio and video recordings were aligned afterward using video editing software.

The child-friendly videos were then edited to include 30 seconds from each language with maximal turn activity and were wedged between entertaining filler videos (e.g., running puppies, singing muppets, flying bugs) for an experimental duration of approximately six minutes long. The order of the non-English videos (videos 2–5) was varied in four versions of the experiment so that no consistent order effects might skew the data. The first and last videos in English (videos 1 and 6) were always kept the same.

2.3 Procedure

Participants were seated in front of an SMI 120Hz corneal reflection eye tracker and a large screen with speakers placed on a table at each side of the screen. The eye-tracker is mounted beneath a flat-panel display; the display is in turn mounted on an ergonomic arm so that it can be positioned at a comfortable height approximately 60cm (an adult arm's length) from the participant. After being seated, participants were told that they would hear videos in a number of different languages. We then asked each participant what languages they could speak. We used a 5-point calibration routine in which participants followed a point on the screen with their eyes. For purposes of engaging children, Elmo (an animated puppet) was used as the calibration image.

In the body of the study, participants watched a six-minute video containing six 30 second dyadic conversations with 15–30 second filler videos between them. The first and last conversations (numbers 1 and 6) were in American English and the intervening conversations (2–5) were recorded in Hebrew, Japanese, German, and Korean. After each conversation, adult participants were asked if they understood any part of the speech to make a second check for any lexical access during the non-English videos.

3 Results and discussion

Child and adult observers in both the English and non-English videos were more likely to keep their eyes on a speaker when that person was speaking rather than when they were silent (Table 1), though they also glanced back at silent participants between 15 and 20% of the time. Children were less likely than adults to keep their eyes on the current speaker while watching the non-English videos, but still showed a reliable difference in gaze to a speaker during speech and during silence. This result indicates that participants were performing basic turn-tracking with their gaze while viewing the stimuli (Kendon, 1967). When point-of-gaze is averaged across the entire recording in this way, there do not appear to be large developmental differences between children and adults in their ability to track the current speaker, though the adults were slightly more consistent.

We next turn to the question of the quick, an-

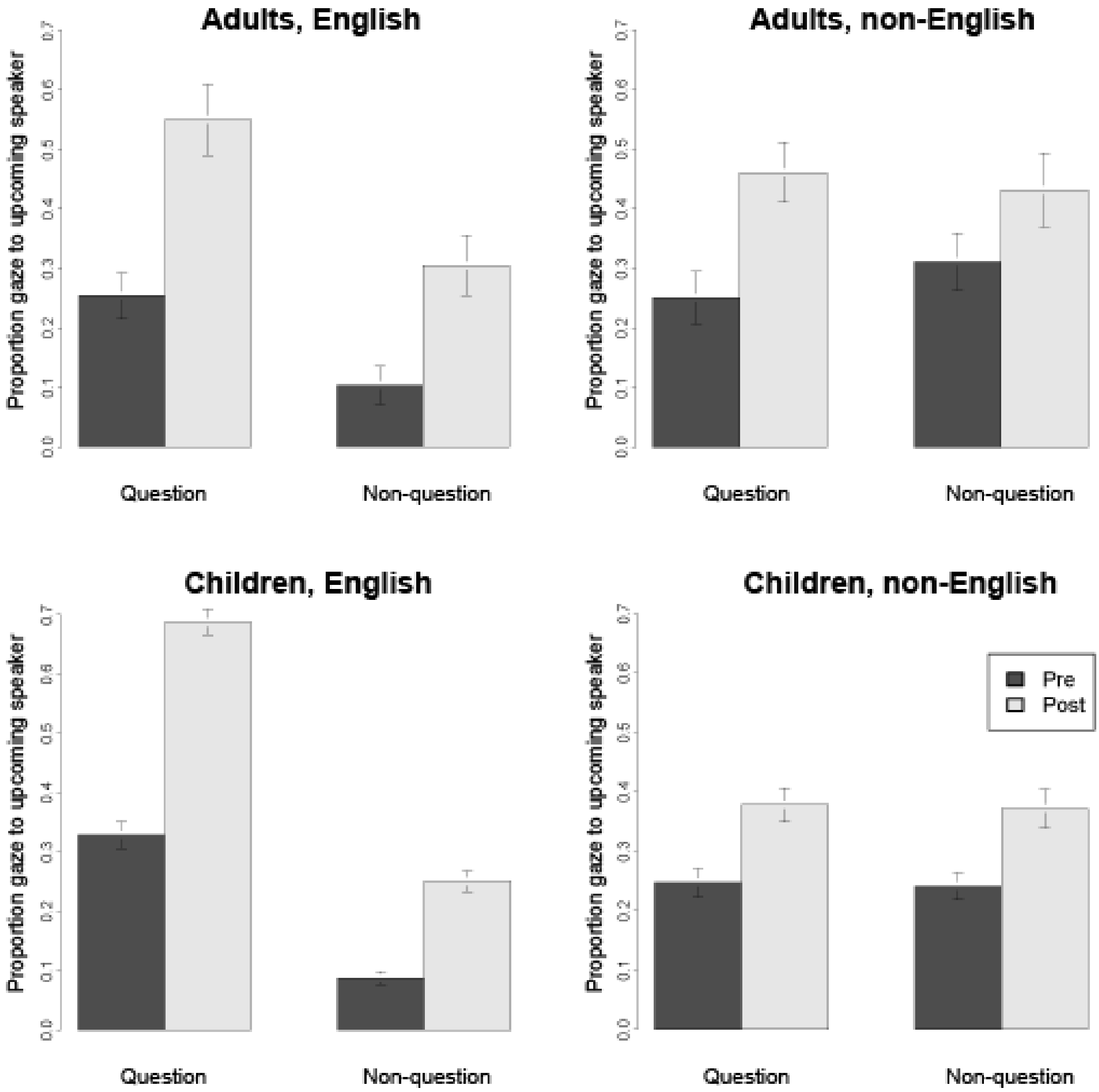


Figure 2: Children and adults' gaze to the upcoming speaker during pre- and post- gap 200 ms windows of speaker switches. Error bars show standard error of the mean across participants.

Group	Language	Current	Non-current
Children	English	0.64	0.17
	Non-Eng	0.48	0.19
Adults	English	0.63	0.16
	Non-Eng	0.61	0.21

Table 1: Average proportion of gaze during speech segments to the current and non-current speaker. Child and adult observers look to the non-current speaker 16–20% of the time the current speaker is talking, and look at neither speaker 18–33% of the time the current speaker is talking. Children watching non-English videos were least likely to be looking at the current speaker during his or her speech.

ticipatory eye-movements around conversational turns observed in previous work (Tice and Henetz, 2011). We test for the presence of turn-end anticipation by measuring shifts in gaze near the inter-speaker gap. Using the average direction of gaze (between previous and upcoming speakers), we compare the 200 ms window prior to the onset of an inter-speaker gap and the 200 ms window following the offset of that inter-speaker gap. Since it would take adults and children at least 200 ms to plan an eye movement, any significant shift in gaze during the 200 ms post-gap window indicates a movement planned prior to the onset of speech by the second speaker. Using this comparison, we find that while viewing English and non-English stimuli, participants tend to anticipate upcoming turn-end boundaries such that they spontaneously shift between the current and previous speaker before the previous speaker has the opportunity to begin his or her response (Figure 1).

Speaker exchanges in the non-English videos that sounded similar to English question-answer adjacency pairs[1] were coded as "questions" for the analysis. We find that both adult and child observers show divergent performance on question and non-question exchanges during all of the videos. Though their gaze begins to shifts dramatically in nearly every case across the pre- and post- gap windows, participants show an advantage for question-answer pairs such that they are more likely to shift earlier on and already be looking at the answerer when he or she begins to speak (Figure 2).

This behavior indicates spontaneous response anticipation during online processing of the stimuli. The average inter-speaker gap across languages and exchange types was 335 ms. The average for questions across languages was 319 ms and 350 ms for non-questions, though the stimuli contain many cases of sub-200 ms inter-speaker gaps. This means that listeners may still rely on a turn-end pause in some cases. However, if participants were universally reacting to silence, we would not expect the earlier switch in question-answer pairs. More generally, the pattern of reliable performance even with inter-speaker gaps shorter than 200 ms suggests that participants make use of cues that are present in the signal prior to the turn-end silence.

We fit two separate linear mixed-effects models (Gelman and Hill, 2007) to participants' average gaze direction at pre- and post- gap windows: one model for adult data and another for the child data. We used a maximal random effects structure to control for variability between participants on the variables of interest. Model coefficients suggest that the advantage for questions over non-questions was significant or nearly significant for both children and adults (t=-7.03 and -1.76, respectively). For children, there was also a significant effect of language group (English vs. non-English, t=-9.29) and a significant interaction between language group and turn type (question vs. non-question, t=6.27). The effect of language group was also nearly significant in the adult data (t=-1.77), and there was no interaction between language group and turn type.

These statistical results suggest that adults were able to integrate non-native cues in their online turn processing more effectively than children were, providing some guidance for an account of the development of turn-end anticipation. For both age groups, there was a significant effect of turn type: question vs. non-question. There may have been many divergent cues in these cases which led participants to earlier and more successful anticipation in the presence of questions. However, since the determination of what counts as a "question" in the non-English videos mainly relied on prosodic similarity to English questions, we have reason to believe that it is precisely because speakers rely on intonational

[1] Judgments were made by the first author primarily based on auditory information, including but not limited to a rising intonation. This judgment is meant to represent which switches the participants were most likely to think were questions in the non-English videos.

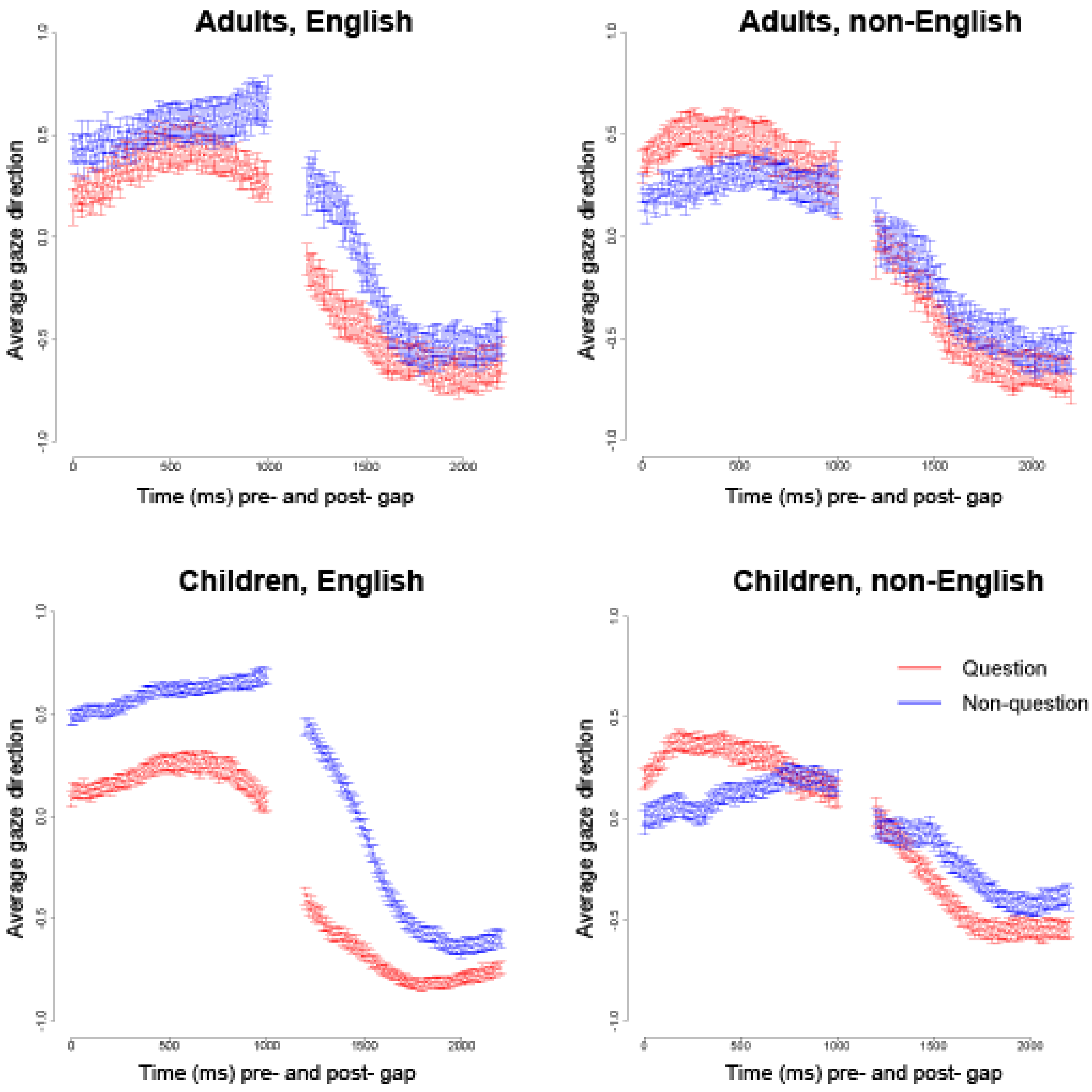


Figure 3: Children and adults' trajectory of gaze over the preceding and following 1-second window of inter-speaker gaps for questions and non-questions in English and non-English videos. Error bars show standard error of the mean across participants.

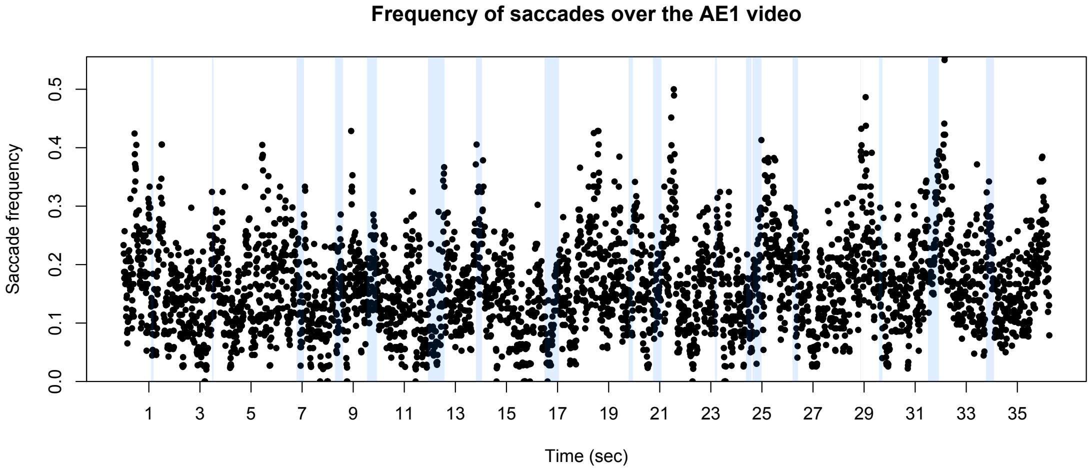


Figure 1: Frequency of saccades over the time course of one video in English. Vertical bars in blue indicate inter-speaker gaps.

Predictor	β	SE	t
Children			
Switches (Non-Question)	-0.31	0.04	-7.03
Lg group (Non-English)	-0.43	0.05	-9.29
Switches x Lg group	0.33	0.05	6.27
Adults			
Switches (Non-Question)	-0.17	0.1	-1.76
Lg group (Non-English)	-0.16	0.09	-1.77
Switches x Lg group	0.005	0.07	0.07

Table 2: Average direction of gaze to the each speaker while he or she is speaking and silent. 1 = looking exclusively at the current speaker and -1 = exclusively at the non-current speaker.

information that they show this advantage. Thus it would be inaccurate to characterize online turn-processing as solely dependent on lexical information. Rather, participants perform remarkably well when no lexical information is present at all.

Consistent with our previous work, the current results provide us with further empirical evidence for spontaneous anticipation of turn-end boundaries. Our results were calculated for without distinction between fixations, long movements, and saccades because of the frequent sampling of the tracker and our decision to analyze anticipation by averaging over pre- and post- gap windows. The anticipatory looking behavior we observed is unlikely to be due to continuous gaze shifting during the video, since saccades show spiked increases only near potential turn boundaries, not between. For example, a time-course rendering of eye-tracking data from one representative video of English conversation shows a considerable spike in saccades prior to turn gaps (Figure 3). Thus, we do not believe that random shifting accounts for our results.

Because each non-English language in this experiment is represented by a single stimulus, we cannot compute reliable across-language differences for each language. Since some of the languages have more overlap in linguistic structure with English, gaze behavior may be significantly better on these items. For example, English speakers can make predictions about the strength of upcoming Swedish prosodic boundaries nearly as well as Swedish speakers do, but Chinese speakers are at a disadvantage in the same task (Carlson et al., 2005). A follow-up study of our work using eye tracking with multiple items from each language would enable us to check for effects of linguistic similarity in languages that the participants do not actually speak.

Finally, in the current study we did not include a baseline condition with no linguistic information at all. Tice & Henetz (2011) found that successful gaze anticipation relies on the presence of linguistic information for English. But, we have no direct comparison of gaze behavior in conditions without any linguistic information and with linguistic information in a language the participants don't speak. This must be added in future work.

4 Conclusion

Children and adults track the current speaker with their gaze. They also spontaneously make anticipatory looks to upcoming speakers at speaker exchanges, indexing their online processing of turn-structure (their anticipation of an ongoing turn's end and the beginning of a responder's turn). Their anticipatory gaze is stronger when prosodic and other non-lexical cues suggest question status (e.g., ending in a high-rise terminal).

Even without lexical information, we track turns as they unfold. Participants not only continued to track current speakers during non-English videos, they showed an advantage for question-type turns over non-question-type turns. A model of how we manage to take turns on time must account for prosodic and other non-lexical information.

We found that adults and children performed almost equally well, with the exception that children had more difficulty maintaining speaker tracking and anticipation during the non-English videos. This may in part be due to their uninhibited lack of interest which resulted in more variable looking patterns than well-behaved adults. Investigation of this possibility will require more data from both age groups and denser developmental data. Children's success in predicting turn-end boundaries and tracking the current speaker suggests that they master this skill early on. It therefore seems likely that their delays in responding to questions (Garvey and Berninger, 1981; Casillas et al., in preparation) has more to do with formulating a response than anticipating when to come in.

In the present study we used recordings of non-English languages to test for turn-processing success when lexical information is not present. Though the non-lexical stimuli are highly naturalistic, they do not directly test which *English* cues English speakers use. There is a significant effect of language group for child participants and a similar, but non-significant effect for adults, suggesting that we can most accurately measure turn-processing performance in English by using English stimuli. To perform the appropriate experiment, we must create phonetically-manipulated stimuli to control for turn-end linguistic cues including prosody and lexical information. We plan to run this follow-up study to compare how performance changes with carefully controlled, but less naturalistic stimuli.

Until recently, we did not have any experimental evidence of turn-end anticipation. But, in the past few years at least two studies have demonstrated that turn-end prediction is a measurable behavior (de Ruiter et al., 2006; Tice and Henetz, 2011). The present study is the first to show evidence that we spontaneously predict turn-end boundaries when attending to languages that we do not speak. This result tells us that the ability to predict upcoming turn-end boundaries is not reliant on lexical information alone; rather, we spontaneously apply (even non-native) prosodic and non-verbal information to continue tracking upcoming turn junctures accurately. Taking all of the experimental work on turn-end anticipation together, our turn processing mechanism is best characterized as a flexible one which makes use of the information available to it in the current conversational environment. These findings indicate that further experimental work will be able to distinguish what cues are attended to as speech unfolds and prediction takes place under different conditions.

5 Acknowledgements

This work is supported by an NSF GRF to M. Casillas. We thank the children, teachers, and directors at the preschool where the child participants were tested. We also thank Tania Henetz, Eve V. Clark, and Herb Clark for their invaluable input.

References

J.B. Bavelas, L. Coates, and T. Johnson. 2002. Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580.

R. Carlson, J. Hirschberg, and M. Swerts. 2005. Cues to upcoming swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, (46):326–333.

M. Casillas, S. Bobb, and E. V. Clark. in preparation. Turn-taking, timing, and access in early language acquisition.

J. Caspers. 2003. Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, 31(2):251–276.

J.-P. de Ruiter, H. Mitterer, and N. J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515.

C.E. Ford and S.A. Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184.

C. Garvey and G. Berninger. 1981. Timing and turn taking in children's conversations 1. *Discourse Processes*, 4(1):27–57.

A. Gelman and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*, volume 1. Cambridge University Press New York.

M. Heldner and J. Edlund. 2010. Pauses, gaps, and overlaps in conversations. *Journal of Phonetics*, (38):555–568.

M. Heldner, J. Edlund, and R. Carlson. 2006. Interruption impossible. In G. Bruce and M. Horne, editors, *Nordic Prosody: Proceedings of the IXth onference*, pages 225–233. Frankfurt am Main: Peter Lang.

A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26(1):22–63.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

T. Stivers, N.J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J.P. De Ruiter, K.E. Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

M. Tice and T. Henetz. 2011. Turn-boundary projection: Looking ahead. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*.

French questioning declaratives: a corpus study

Anne Abeillé, Benoît Crabbé*, Danièle Godard Jean-Marie Marandin
LLF, CNRS, Université Paris Diderot
* Alpage, Université Paris Diderot, Inria

Abstract
French has three types of root polar questions: with *est-ce-que*, with an inverted clitic, or with a declarative clause. We provide semantic and pragmatic properties that set questioning declarative clauses (Q-decl) apart from the two other types, showing that their content is a proposition (like with ordinary declaratives) rather than a question. We provide data from a radio talk show, which illustrate their use in conversation, in comparison with that of interrogatives. They are not regular queries but constitute an appeal to the addressee to endorse the proposition. Using logistic regression models, we point out which factors favor the use of Q-decl over the other types.

1. Introduction

It is usually said that French has three types of clauses that are used for expressing yes-no root questions (Borillo 1978, Mosegaard-Hansen 2001, Marandin 2005). They are illustrated in (1).

(1) a. Est-ce que Paul sera là ?
 'will Paul be there'
 b. Paul sera-t-il là ?
 c. Paul sera là ?

In (1a), the clause is introduced by the interrogative complementizer *est-ce que* ; (1b) is characterized by the presence of a verbal suffix identical to the subject clitic ; the syntactic form of (1c) is identical to that of a declarative clause. It is also usually said that (1c) can be distinguished from an asserted declarative clause by a rising intonation. However, we will leave intonation aside, because the correlation between rising / falling intonation and clause type is not one-to-one. Although most declarative clauses used fro questioning have a rising intonation, this is not true of all, and some such clauses must be recognized on a contextual basis (Mosegaard-Hansen 2001, Beyssade-Marandin 2006).

Some languages do not distinguish between root declaratives and interrogatives on a syntactic basis, while others do. In some, the form used for polar interrogatives is also found in other constructions (this is the case in English, for instance, the subject auxiliary construction being part of a number of constructions, such as conditionals, Fillmore 1999), while a form dedicated for interrogatives is found in other languages. French distinguishes formally between root declaratives and interrogatives, while at the same time using declaratives to ask yes-no questions, and uses both a form dedicated to interrogatives (the complementizer *est-ce que*) and a form also found in different constructions (the verb with a verbal suffix appears in conditionals, with certain initial adverbs etc.). Hence the three forms in (1): the interrogative with *est-ce-que* (est-ce-que-cl) (1a), the inverted clause (inv-cl) (1b), and the declarative clause with a questioning use (Q-decl-cl) (1c).

When different forms are found with similar uses, the natural question is to ask whether they differ, and how. Following Ginzburg and Sag 2000, we distinguish between the semantic type of the clause and its illocutionary import, and, more generally, its uses. Thus,

– *declarative, interrogative* refer to syntactic forms and properties;

– *proposition* and *question* are the semantic types (or content) of declaratives and interrogatives, respectively;

- *assertion, query* refer to the speech act. A speech act is typically (but only typically) associated with a clause type (form + content).

Est-ce-que-cl, *inv-cl* and *Q-decl-cl* are constructions, that is, associations of syntactic, semantic and pragmatic properties.

The plurality of forms in (1) thus raises two different questions: (i) do the forms in (1) differ semantically? More precisely, is the content of a Q-decl-cl a proposition (as expected from its form) or a question (which would be in keeping with what is presented as its use)? (ii) does a Q-decl-cl have the same uses as the interrogatives in (1a,b)? The first question has been raised for English (Gunlogson 2003, Šafářová 2007). Although she does not clearly distinguish between content and use, Gunlogson's proposal includes the idea that the content of a Q-decl-cl is a proposition. The second question is raised for English by the same authors (in particular), but crucially for French by Mosegaard-Hansen (2001). For Gunlogson, it is the intonation which is responsible for allowing the use of a proposition as if it were a question (hence the term rising declarative); crucially, intonation removes the speaker's commitment which is necessary for an assertion. Šafářová (2007) is specially interested in characterizing the contexts appropriate for Q-decl-cl. She proposes three types of contexts (↑ indicates rising intonation, and S the speaker):

– Speaker and addressee are not previously committed to the content of the clause: questions based on contextual evidence (*You're leaving for vacation today* ↑) or try-out statements (S1. *John has to leave early.* S2. *He'll miss the party then* ↑)

– Utterances associated with Speaker's commitment only, whether they introduce new information (*You remember X?),* or check the addressee's approval (S1. *How did you like the movie?* S2. *I thought it was good* ↑)

– Previous addressee only is committed (re-prise or echo questions) (S1. *That copier is broken* S2. *It is* ↑ *Thanks, I'll use a different one*).

Mosegaard-Hansen (M-H) is also mainly concerned with interactional factors favoring one or the other of the three forms in (1). She proposes that two factors are crucial in the choice of a Q-decl-cl, the accessibility of information, and the participation structure of the dialogical situation. The first factor is that Q-decl-cl tend to be about events that are known to the addressee and not to the speaker, while the interrogatives in (1a,b) tend to be about other types of events (e.g. known to both speaker and addressee, to all participants, or on topics known to be disputable). The second factor is that a simple structure favors Q-decl-cl over (1a,b), where a structure is as simple as possible if the roles of speaker and person responsible for the content of the utterance are endorsed by the same person, and similarly for the addressee and audience.

We address both issues. First, we show that the semantic type of Q-decl-cl is a proposition rather than a question. Second, on the basis of previous observations as well as on our intuitions, we make some hypotheses about the illocutionary import of Q-decl-cl. As suggested in M-H (2001), and worked out in Beyssade & Marandin (2006), Farkas & Bruce (2010), Ginzburg (2012), we assume that speech acts are analyzed along two dimensions: the call on addressee (the uptake that the utterance projects for the addressee) and the commitment of the speaker. We test those hypotheses on a corpus (EPAC, Bazillon *et al.* 2011), with quantitative observations, which we check with a statistical model. This corpus is a radio talk show, between a host, callers and experts who answer the callers' questions. We show that the content of a Q-declarative is indeed a proposition (like that of an ordinary declarative. In addition, its use is neither that of an ordinary declarative, nor exactly that of a query. In fact, the speaker hands the content of the clause over to the addressee in order for him/her to take responsibility.

2. French questioning declaratives are propositional

We show that Q-declaratives are indeed declaratives, that is, their content is a proposition rather than a question (like interrogatives). We illustrate our points with attested data, in particular from our corpus.

2.1. Use of the predicate *question*

A Q-declarative is not natural for elaborating the noun *question* (Gunlogson 2003). Examples (2) and (3) propose sentences with an *est-ce-que*-cl and an inv-cl. Q-declaratives cannot be substituted for the interrogatives here. Note that the interrogative clause is the complement of the identity verb; it has the form of a root interrogative because it is used a a quote.

(2) a. mais **la question fondamentale est** : **est-ce qu'** on peut faire des élections libres aujourd'hui, accessibles à tout le monde ? (EPAC file 0813)
'but the fundamental question is, is it possible to organize free elections nowadays, open to everybody'
b. #la question fondamentale est : on peut faire des élections libres aujourd'hui ?

(3) a. **ma question c'est** euh l'enseignement des langues **importe-t-il** à l'éducation nationale ? (EPAC file 0902)
'my question, it is euh is language teaching important for the ministry of Education ?'
b. #ma question c'est euh l'enseignement des langues importe à l'éducation nationale ?

However, the interpretation of such data is not completely straightforward: the noun *question* itself can refer to the act as well as the content of the clause, and can also be the equivalent of 'issue'.

2.2. Polarity subjunctive

Some predicates (verbs of communication, propositional attitudes), which select indicative complement clauses, may, in certain non positive environments, interrogatives among them, take a subjunctive clause. We illustrate the fact with an inverted verb (mood alternation is possible, but less frequent, with *est-ce que,* Huot 1986). On the other hand, Q-declaratives do not license the subjunctive.

(4) a. Vous n'avez pas peur de la mort. Et moi, croyez-vous que je la crains$_{\mathrm{IND}}$ /craigne$_{\mathrm{SUBJ}}$?
'You are not afraid of death. Do you think that I am afraid of it? (J. d'Ormesson, *Le bonheur à San Miniato*, 1987, p. 225)
b. Vous n'avez pas peur de la mort. Et moi, vous croyez que je la crains$_{\mathrm{IND}}$ /*craigne$_{\mathrm{SUBJ}}$?

Broadly, in a semantico-pragmatic approach, the subjunctive mood is motivated in contexts where the interpretation requires taking into account alternative situations (to the one described in the sentence) (Farkas 1992, Godard 2012). An interrogative or a negated belief verb creates such an environment, hence the subjunctive in (4a). The fact that Q-declaratives do not license the subjunctive in the complement of such verbs indicates that their interpretation does not create alternative situations, as interrogatives do.

2.3. Coordination with a *wh-* question

It is difficult to conjoin Q-declaratives with *wh*-questions, while this is completely natural with interrogatives (1a,b).

(5) a. **Que** pensent-ils et **est-ce qu'ils** continuent à faire aveuglément confiance au gouvernement de Georges Bush ? (EPAC file 0813)
'what do they think and do they continue to have full confidence in GB's government '
b. *Que pensent-ils et ils continuent à faire aveuglément confiance au gouvernement de Georges Bush ?

(6) a. **pourquoi** reçoit-on cet avis **et doit-on** le rajouter aux revenus imposables ? (EPAC file 0326)
'why do we receive this notice and must we add it to our taxable income'
b. *Pourquoi reçoit-on cet avis et on doit le rajouter aux revenus imposables ?

This is an indication that the content of Q-declaratives is of a different type from that of interrogatives (1a,b), given the general constraint on symmetrical coordination (with *et*), which requires that the conjuncts be of the same semantic type. If Q-declaratives denoted questions, they should be amenable to coordination with other questions.

2.4. Expression of speaker's attitude

Q-declaratives are compatible with expressions of the speaker's epistemic attitude towards the content of the clause. In particular, they are compatible with markers expressing degrees of certainty, for instance *je crois*, *je présume, peut-être*, which are not felicitous in interrogatives (Gunlogson 2003).

(7) a. Xavier dans l'Isère en ligne. Vous êtes médecin aussi je crois ? (EPAC file 0325)
'X. from Isère on the phone. You are also a doctor, I think ?'
b. *Est-ce que vous êtes médecin aussi je crois ?

(8) a. Et ça s'appelle un CLIS sur les initiales, je présume ? (EPAC file 0902)
'It is called a CLIS on the initials, I suppose ?'
b. *Est-ce que ça s'appelle un CLIS sur les initiales je présume ?

(10) a. beaucoup de questions là-dessus ; on va commencer peut-être avec Étienne Boisserie ? (EPAC file 0402)
'many question on this topic. We'll start maybe with EB ?')
b. *Est-ce qu'on va commencer peut-être avec Etienne Boisserie ?

We may conjecture that the speaker's orientation towards querying requires a complete lack of certainty. In this respect, Q-declaratives are closer to assertions than to queries, since modifiers expressing the speaker's uncertainty are natural in declaratives.

2.5. Tags

Q-declaratives are compatible with question tags (*oui, non, hein, c'est ça, n'est-ce pas*), which are impossible with interrogatives (Beyssade & Marandin 2006).

(11) a. C'est une question de fond quand même euh, **hein** ? (EPAC file 0920)
'but it's a fundamental question really, no ?'
b. *Est-ce que c'est une question de fond quand même, hein ?

(12) a. [le seul moyen de trouver du travail en Irak] c'est euh dans les services de sécurité, **non** ? (EPAC file 0813)
'The only way to find work in Irak, it's euh to work for security, no?'
b*Est-ce que le seul moyen de trouver du travail en Irak , c'est dans les services de sécurité, non ?

A plausible interpretation of this contrast is to say that a tag takes as its argument the content of the sentence it modifies, which must be of type proposition.

3. The use of Q-declaratives and interrogatives: observations and hypotheses

3.1. Presentation of the corpus

The first quantitative study available (to our knowledge) is that of M-H (2001). It is based on a mixed corpus (4h35' of recording) comprising everyday dialogues, radio talk shows and one school examination (note that alternative questions and rhetorical questions have been included, which we exclude). Q-declaratives outnumber the two other constructions (see Table 1). But, as the author herself stresses, the genre is a decisive factor: the distribution in radio debates is more balanced. This suggests that the type of activity or the type of move is relevant in the choice of one of the constructions.

	Q-decl-cl	Est-ce-que-cl	Inv-cl	total
In the whole corpus	204 (83%)	36 (14%)	7 (2%)	247
In radio debates	31 (61,5%)	16 (28%)	6 (10,5%)	53

Table1. Distribution of the 3 forms in M-H

Here we use the EPAC corpus Bazillon *et al.* (2011) which is to our knowledge the only existing French corpus annotated for question types. It consists of transcriptions of 35 sessions of a popular daily radio talk show *Le téléphone sonne* (France Inter) which corresponds to a 20h's recording in 2005, each session lasting about 45 minutes, with a single host, a dozen of experts and a dozen of callers. Callers ask questions over the Internet or the phone about an issue of general concern (tax systems, war in Iraq, etc.), the host reads them or gives the floor to the caller, then he asks experts to elaborate on the topic. Interrogatives have been identified and annotated by the authors: 546 root polar questions have been identified, including Q-declaratives, excluding alternative and rhetorical questions. The distribution of the three constructions is almost even in the corpus (Table 2).

est-ce-q-cl	inv-cl	Q-decl-cl	Total
214 (39%)	154 (28%)	178 (31%)	546

Table2. Distribution of the 3 forms in EPAC

We randomly extracted 17 sessions (10 jours), comprising 258 root polar questions, and studied them for several parameters.

3.2. Frequency and register

It is often said that the three forms in (1) differ in frequency and register: Q-declaratives would be the most frequent in everyday conversations, and would, more generally, belong to an informal register, while est-ce-que-cl and inv-cl would belong to a formal register. However, as shown by M-H, all three forms can co-occur within the same setting. They may well differ regarding some aspect of the context, notably whether the speech is planned or not. We compare the numbers according to the role of the speaker in the conversation (host, caller, expert).

3.3. Speech act

Given that the content of a Q-declarative is a proposition, if the speaker were committed to it, the utterance would simply correspond to an assertion. But this is not the case. Responses with factive adjectives (*Génial*, 'great') are appropriate for assertions, but not for queries (15a,b) or Q-declaratives. Assuming that these adjectives require the content to be part of the set of facts in the common ground (Fernàndez & Ginzburg 2002), if the speaker is not committed to the content of the Q-declarative, the addressee cannot treat it as belonging to the common ground.

(13) a. S1 Il va neiger demain.
'it is going to snow tomorrow'
S2 Super / Génial. ('great')
b. S1 Est-ce qu'il va neiger demain ?
'Is it going to snow tomorrow?'
S2 #Super / #Génial.
c. S1 Il va neiger demain ?
S2 #Super / #Génial.

Note that these responses are appropriate if the Q-decl-cl is interpreted as an indirect proposal (which implies speaker's commitment):

(14) S1 Je fais des lasagnes pour ce soir ?
'I'll make lasagna to-night ?'
S2 Super. ('great')

If the speaker is not committed to the proposition, then the call on addressee must resemble that of a query. Since queries require an explicit response, we compare the numbers of reactions to the different forms. We also look at the kind of response, since a negative response is more natural with an interrogative than with a declarative.

3.4. Relations between participants

As shown by M-H, Q-decl-cl are favored when the addressee is more authoritative on the topic than the speaker. This goes well with the lack of speaker's commitment to the proposition: the speaker appeals to the addressee to take responsibility for the proposition. We look at the number of subjects in the 2nd person, and the presence of a vocative, comparing again Q-decl-cl and the other forms.

Appeal to the addressee is unsufficient to characterize the speaker's attitude. The proposition seems to be presented in such a way that the speaker is invited to agree. To test the hypothesis, we look at the proportion of confirmations and rejections to Q-decl-cl, as well as to the proportion of question tags.

This hypothesis is supported by the fact that Q-declaratives are difficult outside conversational exchanges, as in written questionnaires, where the participants are unknown to each other (the interrogatives belong to free indirect speech, hence the form of a direct root interrogative, with a subject corresponding to the one who asks the question).

(15) a. [about filling tax forms] il faut cocher avec soin les cases relatives au temps de travail: est-ce qu'on est en temps plein ? est-ce qu'on est à temps partiel ? (EPAC file 0326)
'you have to carefully fill the boxes concerning the duration of your employment: do you have a full time job? do you have a part time job?'
b. il faut cocher avec soin les cases relatives au temps de travail: #on est à temps plein ? # on est à temps partiel ?

3.5. Role in conversation

Conversational roles of queries are diverse: topic generation, topic shift, conversation management (giving turns, questioning about who or where the speaker is, whether one hears or understands the question). In our corpus, the main roles for Q-decl are conversation management and topic management.

Q-declaratives may refer to the ongoing conversation more easily than the other types. With epistemic and communication verbs (15a), they may have a null object interpreted as referring to the content just uttered (as in *Tu vois ?* 'You see', *Tu comprends* ? 'you understand'). They are mostly in the present tense, but with modal verbs, they may be in the imperfect (with 2d person subject), for queries referring to the ongoing conversation (15b):

(15)a. Host to expert: vous confirmez, monsieur Doudrich ? (EPAC file 0920)
'You confirm, Mr D. ?'
b . Host: Aude Hapiot, vous vouliez ajouter un commentaire ? (EPAC file 0402)
'A.H., you wanted to add a comment ?'

When used to monitor the cooperation between participants, they may be used as checking or reprise queries, and, more generally, to prevent conversation lapses. In our corpus, the host uses Q-decl to introduce the participants (16), to check their attention and readiness to take the floor (17), (15b), to propose the floor to a participant (mainly the experts) (18). They also serve to make explicit who is speaking to the (absent) audience.

(16) Host to caller : vous êtes je crois à Issy-les Moulineaux Fatiha ? bonsoir !
Caller : oui bonsoir
'you are in I-l-M, I think Fatiha? Good evening' – 'yes, good evening'

(17) Host to expert [on the phone]: vous avez entendu la question de Xavier ? vous l'avez compris ? (EPAC file 0813)

'you heard X's question? You understood it?'

(18) Host to expert : Élisabeth Dupoirier, vous êtes euh (), vous partagez cette analyse?
'E. D., you are euh you agree with this analysis?

Q-declaratives in (17) should be compared with (19a,b): when there is a real trouble on the line, and hence the possibility of an unplanned situation, the host systematically uses inv-cl or est-ce-que. On the other hand, when everything is under control and routinely unfolding, he uses Q-declaratives. As for (18), the host gives the floor by anticipating the expert's opinion about the topic. He uses interrogatives when he anticipates the expert might disagree with the caller's or his own opinion or orientation (19c).

(19) a. Host to caller: Frédéric ne nous entend pas. Frédéric, est-ce que vous m'entendez bien là, Frédéric ? Frédéric ? non, visiblement il ne nous entend pas (EPAC file0825)
'F. does not hear us. F. , do you hear me F. ? F.? no, clearly, he does not hear us'
b. Host to caller: êtes-vous toujours là, madame ? non, elle n'est plus là (EPAC file0326)
'are you still there, madame ? no, she is no longer connected'
c. Host to expert: est-ce qu'on peut en dire un mot ? […] monsieur Dubois ?
[no answer] (EPAC file 0920) 'can we say a few words about it, Mr D.'

Q-declaratives may also be used for topic management. The formulaic *Tu sais, Tu as vu ce qui est arrivé à X* (you know, you've seen what happened to X…) are used to introduce a new topic by a speaker who pretends that the addressee is already informed. In our corpus, the host uses Q-declaratives to reformulate a question (20), to reprise an assertion literally or quasi literally (21), or to refocus on the current discourse topic. Dislocated constructions (left (20) and right dislocation (21)) are typically used in the last case:

(20) Host: **votre question** Stéphane en fait c'est sur les prix tout simplement ?
Caller: sur les prix, sur la qualité aussi (EPAC file 0816)
'your question, S., it concerns really just prices?' '– it's about prices, about quality too'

(21) [about a new cancer hot line] Host: il est d'ores et déjà opérationnel **ce numéro** ?
'it's already operational, this number?'
Expert: oui il est ouvert depuis lundi (EPAC file 0325)
'yes, it's been in operation since Monday'

We annotate our corpus with two variables: conversation management (yes/no) and dislocation (yes/no) for topic management.

4. Usage properties of the Q-Declaratives

As it stands, the EPAC corpus is already annotated for a question type variable **Q-type** whose values are (Est-ce, Decl, Inv-cl) and a speaker identity variable **SpkI** being an enumeration of the 90 different speakers. As can be seen from a first observation of the data, the overall distribution of question type is roughly uniform (see Table 2, Figure 1).

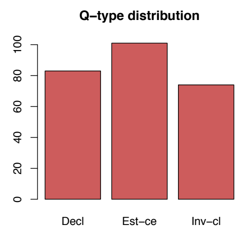


Figure 1: Q-decl distribution

We further annotated the data with variables identified to be of interest for our study. These are :

- **SpkS,** indicates the status of the speaker in dialogue (Host, Caller, Expert).
- **Subj**, the person of the subject (2^{nd}, other)[1]
- **Voc**, the absence or presence of a vocative (yes,no)
- **Disl**, the absence or presence of a dislocated phrase (yes,no).
- **Polarity** of the clause (+, -)
- **Tense**, the clause tense (present, other)
- **ConvM**, the conversational management role (yes,no)
- **Question tags** such as *non, oui, hein, c'est ça*…(yes,no)
- **Resp**, encodes the answer type (confirmation, rejection, no-answer)

[1] 2nd person subjects are mostly polite *vous*.

We also took into consideration a variable which is a potential cause of some idiosyncratic random variation, that is the speaker identity **SpkI** (since we have 90 different speakers).
In order to identify the properties specific to the Q-declaratives, we fit three mixed effect logistic regression models, one for each type of clause and then compare the significant factors for each model. Each such model attempts to predict a binary variable given all the above mentioned predictors on the full dataset.

4.1 The statistical models

The Q-declarative model attempts to predict the positive outcomes of a binary variable Q_decl (yes,no) given the vector of variables **X** including all the above mentioned predictors.
We first observe that **SpkS=Caller** is a categorical predictor for declarative questions since no Caller ever utters a Q-decl-cl as outlined in Table 3:

	SpkS=Caller	SpkS=Expert	SpkS=Host
Q-decl=yes	0	7	76
Q-decl=no	90	29	56

Table 3: Joint distribution of Q-decl and SpkS

To avoid numerical instability, we therefore recoded the tri-valued **SpkS** variable as a two valued **SpkS'** variable (Host, Other).
Another perfect predictor is the question tag variable, which is only found in Q-decl (Table 4). As a consequence we did not include it in our models.

	Q-tag=no	Q-tag=yes
Q-decl=no	175	0
Q-decl=yes	66	17

Table 4: Joint distribution of Q-decl and Q-tag

All the remaining variables are set as fixed effects except the variable **SpkI** set as random effect, yielding the model :

$$P(Q\_decl = yes \mid \mathbf{X}) = \frac{e^{\alpha + SpkI_i + \beta\mathbf{X}}}{1 + e^{\alpha + SpkI_i + \beta\mathbf{X}}}$$

The full model is then reduced by model comparison where we can remove the polarity, subject, vocative and tense variables (likelihood ratio test : χ^2 p-val=0.24). The random effect has an almost null variance and is removed as well[2].

	Estimate	Std. Err	z value	Pr(>|z|)	
(Intercept)	-0.1965	0.3462	-0.568	0.5702	
speaker = Host	2.5212	0.4950	-5.093	3.52e-07	***
conv.mgt= yes	2.0396	0.4570	4.463	8.07e-06	***
Resp = rej	-1.1996	0.5206	-2.304	0.0212	*
Resp = none	-1.2657	0.4887	-2.590	0.0096	**
Disloc = yes	1.4462	0.6552	2.207	0.0273	*

Model 1**:** Q-Declaratives model

The goodness of fit of the model is satisfactory (accurracy = 84%).
Five out of six remaining effects are significant. First, as shown in Table 3 and Figure 2, the speaker status reveals that the host status clearly favors the use of Q-decl while callers and experts rarely use them:

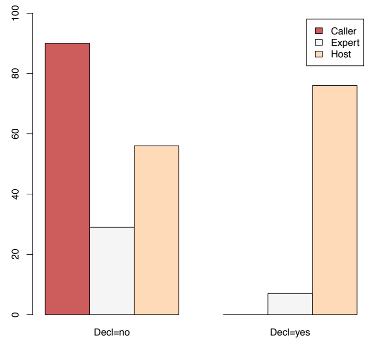


Figure 2: Q-declaratives given **SpkS**

The use of Q-declaratives is also favored by contexts of conversation management as shown in Figure 3 where we can see that more than the half of Q-declaratives are uttered in contexts of conversation management:

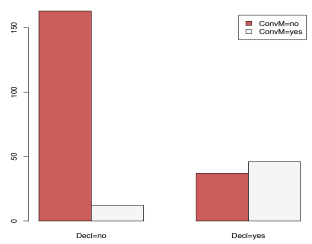


Figure 3 : Q-declaratives given conversation management

[2] We do not have enough data to get proper convergence when fitting a model with all possible interactions among all variables. However we get proper convergence when fitting model 1 augmented with all interactions of order 2. It turns out than none of them are significant. The model with interactions can indeed be reduced back to model 1 without interactions (likelihood ratio test : χ^2 p-val=0.41).

As one might have expected, conversation management and speaker status are strongly associated variables, as shown in Table 5 where we can see that the Host is largely responsible for managing the conversation.

	SpkS=Host	SpkS=Other
ConvM = No	81	119
ConvM = Yes	51	7

Table 5: Non independence of **ConvM** and **SpkS** variables ($\chi^2 = 5.187\ 10^{-10}$)

Third, regarding responses, we see in Figure 4 that a Q-decl strongly favors a confirmation answer whereas other types don't.

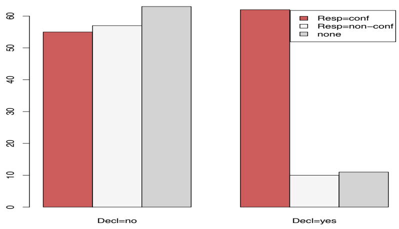


Figure 4: Q-declaratives given Response types

Regarding dislocations, matters are less clear. Although the variable is significant and cannot be removed from the model (likelihood ratio test : χ^2 p=0.02), the only thing we can observe is that the proportion of dislocations in the context of Q-decl is approximately twice the proportion of dislocations in the context of an interrogative clause as shown in Table 6:

	Disl=no	Disl=yes
Decl=no	166	9
Decl=yes	74	9

Table 6: Joint distribution of Dislocation and Declarative variables

In order to contrast the use of Q-decl with other types of questions, we also fitted two additional models, one for *Est-ce que*-cl and another one for Inv-cl.

We used a similar model and protocol to predict the *Est-ce que*-cl on the same data set using as fixed effect predictors the same variables as above and the Speaker Identity as a random effect:

$$P(Est-ce-que = yes \mid \mathbf{X}) = \frac{e^{\alpha + SpkI_i + \beta\mathbf{X}}}{1 + e^{\alpha + SpkI_i + \beta\mathbf{X}}}$$

We also reduce the full model by removing the non significant factors : polarity, vocative, conv.management, dislocation, tag and tense variables (likelihood ratio test, χ^2=0.16).

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.7594	0.7773	-4.837	1.32e-06	***
speaker = Host	-1.4150	0.5892	2.401	0.016333	*
subject =2	-1.8603	0.5642	3.297	0.000977	***
Resp=rej	0.4009	0.4263	0.940	0.346997	
Resp=none	1.6056	0.4125	3.893	9.91e-05	***

Model 2: Est-ce que-clause

Finally, our third model amounts to predict a binary *inv-cl* variable on the same data set with the same protocol:

$$P(inv-cl = yes \mid \mathbf{X}) = \frac{e^{\alpha + SpkI_i + \beta\mathbf{X}}}{1 + e^{\alpha + SpkI_i + \beta\mathbf{X}}}$$

By model reduction, we remove the speaker, polarity, vocative, answer and tag variables (likelihood ratio test, χ^2=0.37).

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.6916	1.1981	-0.577	0.56375	
Conv.mgmt= yes	-1.0792	0.5973	-1.807	0.07079	.
Dislocation= yes	-3.7507	2.2490	-1.668	0.09537	.
Resp = rej	1.2634	0.5427	2.328	0.01992	*
Resp = none	0.2708	0.5738	0.472	0.63693	
Tense = pres	2.1575	0.8483	2.543	0.01098	*

Model 3: Inv-clause

4.2 Synthesis

In order to identify the characteristic properties of Q-decl, we contrast our different models. In Table 7, we cross each factor with each clause type. Cells are ticked with + (respectively -) when the factor is significant with a positive (respectively negative) coefficient for the question type and with (×) when not significant[3].

	Q-DECL	EST-CE QUE	INV-CL
SpkS=Host	+	-	×
ConvM	+	×	-
Resp=rej	-	+	+
Resp=none	-	+	×
Disl=yes	+	×	-
Q-tag=yes	+	×	×
Subj=2	×	-	×
Tense=pres	×	×	+

Table 7: A comparison of the 3 models

[3] Although it is a categorical predictor not included in the actual models, we have also included for completeness the question tag variable in the table.

The main observation is that a Q-decl is mainly used in context of conversation management by the Host speaker (Figure 5 and Figure 6), which can be explained since the corpus is a multiparty conversation with certain speakers over the phone: the host is assigning turns, questioning about who or where the addressee is, whether he or she hears or understands the question etc. Q-decl are primarily used for conversation management (57% of the time).

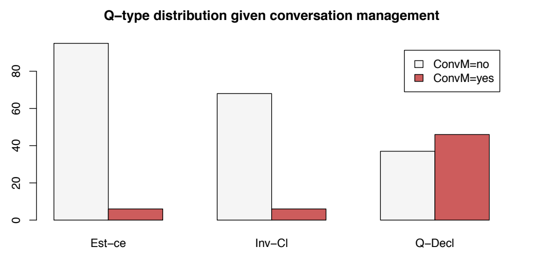


Figure 5: clause-type distribution given conversation management

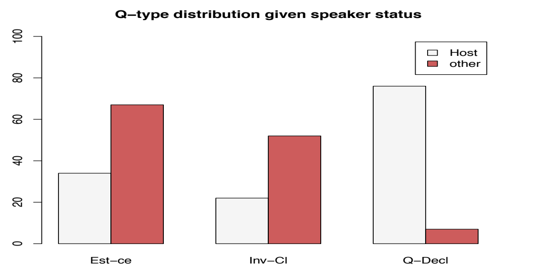


Figure 6: clause-type distribution given **SpkS**

As we have already seen in Table 4, the Q-declarative is the only clause type which is used with question tags.

Q-decl mostly receive a confirmation response (74%) while other types exhibit a more balanced distribution : 28% confirmative responses for est-ce-que-cl, 35 % for inv-cl.

Responses with *oui, non, si* ('yes', 'no') are appropriate both after assertions and after queries. But assertions differ from queries because they do not require an explicit response, viz. an answer or a reply related to the issue raised by the query (although some sort of reaction, possibly non linguistic, seems to be needed). In this respect, Q-declaratives resemble queries: they require an explicit response. In our corpus, 85% of Q-declaratives receive an explicit response.

Confirmation responses are not favored with est-ce-que and inv-cl. While lack of response and rejective response are both factors of est-ce-que and inv-cl models, the lack of an explicit response is only significant with est-ce-que, and rejective response is only significant with inv-cl.

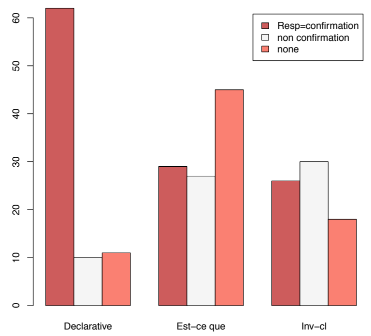


Figure 7: clause-type distribution given Response

4.3. Further explorations

There are two variables for which we expected to get some effects which cannot be observed by our models. The Vocative variable is the first of them. Although it never plays a significant role for predicting a given type of question independently of the others. Nevertheless, the distribution of the vocative variable is strongly dependent on the clause type as can be seen in Figure 9:

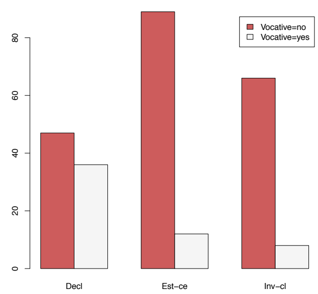


Figure 9: Q-type distribution given Vocative

Proportionally, vocatives are indeed used more in Q-Decl than in other types. This is not a matter of chance, a chi square test of independence between the two binary variables Decl(+,-) and Vocative(+,-) reveals a true association (χ^2; p-val=1.58 10^{-8}).
The other variable that does not show up immediately is the 2d person subject which is more present in proportion in Q-decl than in

other types as shown up by Figure 10. Again there is a strong association between the Decl(+,-) and 2nd person(+,-) variables (χ^2; p-val=3.69 10^{-10}).

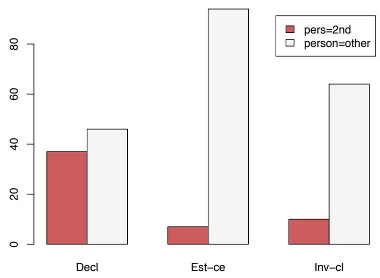


Figure 10: Q-type given 2d pers Subject

The type of response may be associated with the clause polarity. Cooper & Ginzburg 2011 claimed on the basis of a corpus study (BNC) that positive questions tend to be answered positively and negative ones negativeley. It is true that our Q-decl are most of the time positive clauses. However, clause polarity did not appear as significant among our types, nor did we find an association between polarity and clause type (Table 8, (χ^2= 0.12).

	Q-decl	Est-ce que	inv-cl
Polarity=+	78	74	56
Polarity=-	5	27	18

Table 8 : Polarity given clause type

Conclusion

We have shown that, semantically, Q-declaratives are bona fide declaratives (their content is a proposition). Pragmatically, we compare the properties of the three different forms in a radio talk show. The main features that emerge are the following: Q-decl like queries, and unlike assertions, are followed by an explicit response, and they tend to be confirmed. They have two main roles: conversation management and topic management.

It remains to be seen whether these properties hold in other situations. While the first might be correlated (in part) with the status of the speaker who uses them (the host), we expect the second property to be more general. In future work, prosody should be taken into consideration and might provide further distinctions among our 3 types.

Another open question is how our results may extend to other languages (such as Hebrew or modern Greek) which also have more than one type of polar questions, among them Q-declaratives.

Selected references

Bazillon, T., B. Maza, M. Rouvier, F. Béchet, A. Nasr. 2011. Qui êtes-vous ? Catégoriser les questions pour déterminer le rôle des locuteurs dans des conversations orales, Proc. TALN.

Beyssade, C. & J-M. Marandin. 2006. The Speech Act Assignment Problem Revisited: Disentangling Speaker's Commitment from Speaker's Call on Addressee. In O. Bonami & P. Cabredo (eds), Empirical issues in Syntax and Semantics 6, 37-68.

Borillo A. 1978. Structure et valeur énonciative de l'interrogation totale en français. Thèse d'Etat. Université d'Aix-en-Provence.

Cooper, R., & J. Ginzburg. 2011. Negative inquisitiveness and alternatives based negation, in M. Aloni et al. (eds.) Proceedings 18th Amsterdam colloquium.

Huot H. 1986, Le subjonctif dans les complétives : subjectivité et modélisation, in M. Ronat, D. Couquaux (ed) La grammaire modulaire, Paris : Minuit, 81-111.

Farkas, D. & K. Bruce. 2010. On reacting to assertions and polar questions. Journal of semantics 27: 81-118.

Fernàndez R. & Ginzburg J. 2002. A corpus study of Non-sentential utterances in dialogue, *TAL*, 43-2:12-43.

Fillmore, C. 1999. Inversion and constructional inheritance. In G. Webelhuth et al. (eds.), Lexical and Constructional Aspects of Aspects of Linguistic Explanation. Stanford: CSLI Publications, 113–128.

Marandin, J-M. 2005. Formatage de l'information : focus et contexte. In F. Corblin & C. Gardent (eds.), Interpréter en contexte, 31-80, London: Hermès.

Ginzburg, J. & I. A. Sag. 2000. Interrogative investigations, Stanford: CSLI Publications.

Gunlogson, C. 2003. True to form: rising and falling declaratives as questions in English. New York: Routledge.

Mosegaard Hansen, M-B. 2001. Syntax in interaction. Form and function of yes/no interrogatives in spoken standard French, Studies in language, 463-520.

Šafářová, M. 2007. Nuclear rises in update semantics. In Questions, Elsevier.

The Use of Gesture to Communicate about Felt Experiences

Nicola Plant and **Patrick G.T. Healey**
nplant@eecs.qmul.ac.uk ph@eecs.qmul.ac.uk
Queen Mary University of London
Interaction, Media and Communication Research Group
School of Electronic Engineering and Computer Science
London, UK

Abstract

When people describe the character of felt experiences such as a headache they can use their bodies as a resource to help them communicate. It has been proposed that when speakers use gestures, pose and facial expressions to describe an experience their listeners simulate or mimic these cues in order to help them understand the character the experience. We test this model using data from dyadic conversations in a laboratory setting. The results show that listeners do not normally match the expressive gestures that speakers use in describing their experiences and that while speakers gesture more strongly for more negative experiences their listeners do not. Rather than re-creating the speaker's experience through mimicry, listener gestures appear to be used primarily for engaging with the concrete particulars of an experience and not its 'subjective' effects.

1 Introduction

Communicating the character of a felt sensation, such as a pain, seems harder than communicating about, say, the concrete events that caused a pain (although see (Wittgenstein, 1958)). Felt experience can be difficult to articulate and there is no guarantee that it is shared from person to person. Nonetheless, we do have the capacity to empathise with each others experience although the particular mechanisms behind empathy are still disputed and unclear (Preston and de Waal, 2002).

Much of the debate about empathy concerns the in principle possibility (or impossibility) of knowing another's experience. Here we are concerned with the empirical question of how people actually go about trying to communicate their experiences in conversation. In particular, we are interested in the role of gesture, and embodiment more generally, in this process.

Gestures, posture and facial expressions are frequently used in conversation to complement and provide additional information to the accompanying speech (Kendon, 2004; McNeill, Cassell and Levy, 1998). Temporally and semantically coupled with the verbal elements raised in speech, these embodied depictions can potentially provide a more direct representation of the imagistic and embodied aspects of a speaker's message. Gestures can be particularly useful when depicting the visual or spatial domain, due to their own physicality. Gestures, posture and expressions can also provide a more direct expression of an embodied experience through direct displays or demonstrations of aspects of an experience.

In order to obtain a better understanding of how people use gestures, posture and expressions in communicating about their experiences we have collected a corpus of speech, video and body movement data. In this corpus, pairs of participants take turns to describe to each other recalled experiences that invoke significant elements of embodied experience, for example a toothache or a yawn. Here we report an initial investigation of the use of gesture in the expression of these experiences in an interaction. A basic premise of our approach is that it is the use of gesture in the live dynamics of an exchange rather than, for example, in telling a story to a camera or an experimental confederate, that is key to understanding how experience is communicated in practice. It is also clear from the literature that gestures in interaction differ from those produced in monologue (Bavelas and Gerwing, 2011). More specifically, we are interested in how speakers' gestural descriptions of their experience contribute to the listeners' understanding and what *listeners* do in response to demonstrate their understanding.

A key hypothesis about how expressions and gestures contribute to communication is that listener's automatically mimic them in order to 'simulate' or 'reconstruct' the described experience and, thereby, enhance their understanding of what was said. For example, Hatfield et al. (1993) claim there is an automatic and continuous tendency to mimic one another in social interactions. Defining mimicry as a form of synchronisation of posture, facial expression, movement and instrumental behaviour of others including mimicking and synchronising vocal utterances (Hatfield, Capioppo and Rapson, 1993). Similarly, Chartrand and Bargh (1999), drawing on James principle of Ideomotor-action, propose that merely thinking about behaviour increases the tendency to engage in that behaviour. This perception-behaviour link is, they claim, a natural and non-conscious connection between the act of perceiving and the act of behaving, such that perceiving an action being done by another automatically makes one more likely to engage in the same behaviour. They suggest a two-step process for this process of direct environmental causation of social behaviour: (1) automatic perceptual categorisation and interpretation of social behaviour (environment to perception), (2) perceptual activation continuing on to activate corresponding behaviour representation (perception to behaviour). Here, they propose the sequence from perception to behaviour occurs entirely automatically, so should occur even amongst strangers and should occur even in the absence of a reason to do so, such as pursuing an affiliation goal (Chartrand and Bargh, 1999).

Two predictions follow from this model. First, that speaker descriptions of felt experiences accompanied with gestural expressions should cause the listener to produce similar behaviour or gestures. Second, that the strength and form of the listeners responses should match the strength and form of the gestures produced by the speaker. More specifically, the more unpleasant or painful (negative) the experience being described by the speaker the more mimicry produced through the empathetic responses produced by the listener in sympathy with the speaker.

We test these predictions against the behaviour of speakers and listeners in our corpus. In order to do this we first introduce a taxonomy of gesture types used to code participants responses.

1.1 A Taxonomy of Descriptive Gestures

There are many different uses of gesture and many different gesture taxonomies. For current purposes a simple taxonomy is required that allows us to discriminate basic functional differences. A review of literature categorising different types of gestures within an interaction follows, seeking to build a taxonomy of each in which a suitable context for the gestural descriptions of sensation can be placed.

Gesture, for current purposes, includes anything non-verbal[1] that is produced as part of the conversation including hand movements, postures shifts and so on. For a simple taxonomy to categorise gestures describing felt experience, we exclude gestures that form para or meta narrative elements of interaction. Para-narrative elements are where gestures are about or managing the interaction itself rather than the semantic content of what is being communicated. Meta-narrative elements are where gestures are about the speech itself, again not the semantic content or the interaction. What we are interested in are the narrative elements of an interaction, these refer to gestures that depict or are about the content of the speech. For the purposes of our study we focus on 'narrative' or 'topic' gestures (Bavelas, Chovil and Lawrie, 1992) that relate to the content of the speech. These are spontaneous gestures that occur mostly synchronously with speech, also referred to as as physiographic (Efron, 1941) or lexical movements (Krauss, Chen and Chawla, 1996).

2 Method

In order to elicit unscripted accounts of felt experiences we asked people to describe to each other

[1]Note: We use non-verbal as a gloss to mean non-speech communicative actions such as conversational facial displays, gestures and body movements.

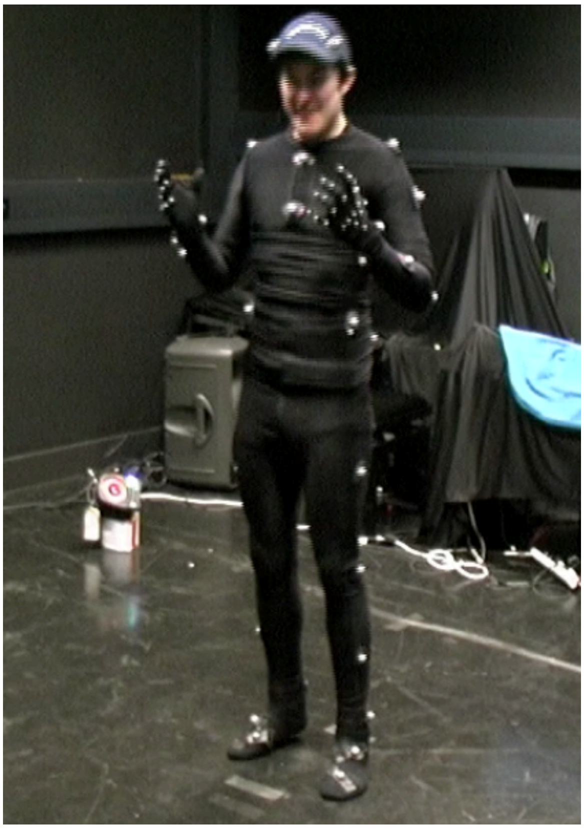


Figure 1: An Intense Throbbing Pain.

recalled experiences such as a toothache or a yawn, that have an embodied element and could provoke empathetic responses.

2.1 Participants

A total of 24 participants were recruited. Participants ages ranged from 18 to 60, consisting of 12 females and 12 males placed in 12 pairs. The aim was to elicit unrehearsed and spontaneous descriptions in an interaction. However, as the study required participants to wear motion capture suits with reflective markers, we were aware of the effect of this on the interaction, such as the assumption that participants movements (and by extension, their gestures) were being studied.

2.2 Materials

The corpus of speech, video and body movement data were captured in the Performance Laboratory at QMUL. Video footage included a full body face on view of each participant for the duration of the study. Motion capture data was also obtained for participant using a Vicon optical marker system although we do not report this data here. A set of cards were placed on a small table next to where the participants stood. Each participant was given a stack of these cards and were asked to take turns selecting one card at a time. There were 8 cards in total per participant, the experiences written on sets of cards were a headache, the taste of a nice meal, a toothache, a stomach ache, a backache, a yawn, laughing out loud and a back massage.

2.3 Procedure

The participants were given written instructions outlining the study procedure. They were asked to recall specific instances of the experiences stated on the cards and talk about them to each other. When it was their turn each participant was to explain the details of a previous experience they have had of the sensation written on the card to their partner for no longer than a 2 minutes. An emphasis was placed on describing how this experience felt or the particular sensation they felt at the time of the experience. On each description the listening participant was encouraged to talk and ask questions at any time, the process was described in the instructions as an exchange. In attempt to allow participants to practise and settle into the irregular nature of having a conversation in the conditions of the lab, the first two experiences that came up in the set of cards in each session were practise experiences that were not analysed, these were a headache and the taste of a nice meal. Aside from the two practise experiences at the beginning of the stack, the cards were shuffled into random order for both participants in each session.

2.4 Coding

For the coding process, the video and audio descriptions of each experience was separated into separate items. As we are interested in topic gesture, these were coded for each item. The participants were labelled cardholder (CH) and non-cardholder (NCH) for each item. On the first pass, any occurrences of physiographic gesture were coded without specifying their nature. On the second pass, topic gestures were separated into three types: Iconic, Metaphoric and Abstract Descriptive, the annotator following the definitions indicated in the taxonomy below. It is important to mention that all topic gestures were coded irrespective if it was the cardholder or non-cardholder that performed it, so both speaker and listener were coded for their gesture and the same definitions were used for each. On each pass, only one camera view was coded at a time, so for example, while coding the cardholders gestures the non-cardholder was not in view.

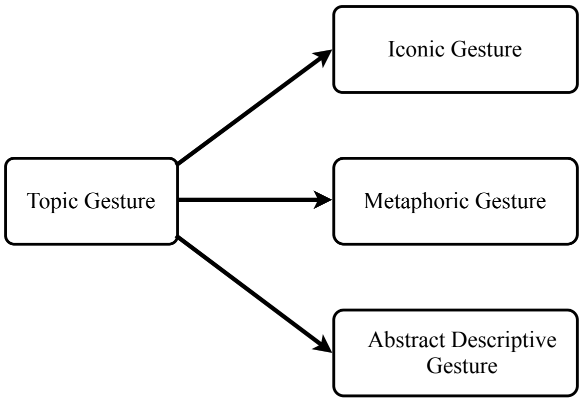


Figure 2: Gesture Taxonomy.

In order to differentiate between narrative gesture types in this study we adopt McNeill's subdivision of spontaneous gesture into Iconic and Metaphoric subcategories (McNeill, Cassell and Levy, 1998) and supplement it with the 'Abstract Descriptive' topic gestures described by (Rowbotham et al., 2011).

Iconic gestures include a depiction that is intrinsic to the content that is being conveyed, for example a gesture describing a ball will depict a characteristic intrinsic to a ball, such as making a fist to represent roundness by making the hand round. Iconic gestures are always a concrete entity or action rather than abstract or analogous. Perhaps more relevant to us is iconic gestures that act out being in pain or the cause of the pain. For example, when describing an experience of stubbing a toe, the speaker may hop around on one foot depicting the outward behaviour caused by stubbing their toe. This mimicry of their behaviour is intrinsic to the original action and so therefore we class as a iconic gesture. They encode a speakers viewpoint on a communicated depiction. Ekman and Friesen (1969) include deictics, spatial, kinetographs, pictographs and rhythmic elements in this category. Beattie (2002) suggests that the content of these gestures to describe action, shape, size, direction, speed, and relative-positions.

Metaphoric gestures: are usually pictorial but unlike iconic gestures, metaphorics depict abstract ideas rather than a concrete object or event. The topic of the metaphor being the abstraction, the vehicle or gestural image being the offered virtual object spatially localised, and the common ground of meaning is where the topic and vehicle are linked in properties, such as physical containers. For example, in a description of a backache an interlocutor describes the most extreme sensation of that particular pain by saying 'and that was the crest of it', while reaching high and pointing to the top point in the gesture space, spatially highest gesture metaphorically signifying the most intense pain. We exclude what Gullberg (2009) calls conventionalised gesture, otherwise known as emblems and symbolic gesture (Ekman and Friesen, 1969; Krauss, Chen and Chawla, 1996). These have a known meaning across a culture and are independent of speech, although can accompany it. Here we exclude as they lack the spontaneous descriptive characteristic that focuses on the quality of the experience.

Absract Descriptives Rowbotham, Holler and Lloyd (2011) make an additional subtype of topic gesture they term abstract descriptive, these gestures describe inner experiences and are categorised by featuring 'imaginist and semantically related to speech but containing information which could not be visually accessed (subjective experience of pain), therefore not iconic according to McNeill". (Rowbotham et al., 2011; Hyden and Peolsson, 2002) These gestures describe the felt sensation without acting it out, and express more than just the location of the sensation. They specifically refer to the inner sensation of the experience, put plainly, what the sensation feels like. For example, when describing the sensation of stubbing ones toe we might depict the pain with a gesture that uses our hands to represent the rhythmic quality of a throbbing pain by mimicking it in the rhythm of our hand movement, or perhaps the intensity intrinsic to the pain would be depicted by the tenseness of our fingers. This type of gesture does not represent an analogy of the felt experience so cannot be a metaphoric gesture but on the other hand cannot be symbolised concretely, as the felt experience is only accessible to the experiencer. This indicates that they lie somewhere on the borders of iconic gesture and metaphoric gesture.

3 Results

We report data for 9 pairs of participants and for four target items: Toothache, Backache, Yawn and Laugh. Excluded were a stomach ache and a back massage, these were excluded because for these particular items, one pair of participants proceeded to talk about another persons experience of the sensation on their card, for example describing their fathers backache or a dogs yawn, this resulted in although unrehearsed and spontaneous description, not a personal account. Also excluded were two sessions where the participants didn't follow the instructions, where both participants talked about one experience that came up on one participants card at the same time, comparing experiences rather than describing their own individual experiences, these were very different interactions to the other sessions.

The overall distribution of different gesture types is reported in Table 1.

Gesture Type	Mean	Std. Deviation	Median	Sum
Abstract Descriptive	1.56	2.090	1.00	184
Iconic	3.58	4.463	2.50	423
Metaphoric	.07	.252	.00	8

Table 1: Overall Occurances of Different Gesture Types

Because Metaphoric gestures were low frequency in these data (0.07 per item, less that 1.3%) they are excluded from the statistical analysis.

The initial inspection of the data for the dependent variables, number of occurrences and duration of gestures, showed a strong positive skew towards zero so a Tweedie distribution was used for the Generalised Estimating Equations analyses reported below.

Item	Mean	Std. Deviation	Median	Sum
Backache	2.88	3.844	2.00	184
Laugh	1.89	2.356	1.00	106
Toothache	3.52	5.106	1.50	197
Yawn	2.00	2.300	1.00	120

Table 2: Average Gesture Use for Each Target Item (Excluding Metaphorics)

The four target items are not directly comparable but were ranked according to Experience Type to reflect the intuitive degree of (un)pleasantness involved. For analysis we ranked them in the variable Experience Type as follows: 1 Laugh, 2 Yawn, 3 Backache, 4 Toothache to provide a ranking from positive to negative experience.

Dependent Variable: Occurrences

	Type III		
Source	Wald Chi-Square	df	Sig.
(Intercept)	33.655	1	.000
Role	49.821	1	.000
Gesture Type	19.700	1	.000
Experience Type	11.499	3	.009
Role * Gesture Type	5.135	1	.023
Role * Experience Type	1.332	3	.722
Gesture Type * Experience Type	2.635	3	.451
Role * Gesture Type * Experience Type	4.042	3	.257

Table 3: Results of GEE Analysis for Frequency of Gestures

Dependent Variable: Average Duration

	Type III		
Source	Wald Chi-Square	df	Sig.
(Intercept)	20.766	1	.000
Role	51.953	1	.000
Gesture Type	1.924	1	.165
Experience Type	5.178	3	.159
Role * Gesture Type	4.485	1	.034
Role * Experience Type	2.961	3	.398
Gesture Type * Experience Type	5.994	3	.112
Role * Gesture Type * Experience Type	3.560	3	.313

Table 4: Results of GEE Analysis for Average Duration of Gestures

In order to check if the order of presentation has an effect on the frequency of gesturing item number was correlated with total number of gestures produced this was not significant (Kendalls Tau-b = 0.03, p (two-tailed) = 0.49.

Two main Generalised Estimating Equation (GEE) analyses were carried out on a) the frequency of occurrence of gestures and b) the duration of gestures under the conditions defined by the three experimental factors: Role of the participant (Cardholder vs. Non-Cardholder), Gesture Type (Iconic vs Abstract Descriptive) and Experience Type (from 1 positive to 4 negative). All two-way and three-way interactions were included. Participant ID was included as a subject factor. Pair identity, Role, Order and Item specified as within-subjects variables.

As Table 3 shows, there were main effects of Role, Gesture Type and Experience Type and a Role X Gesture Type Interaction on the likelihood that a gesture would be produced (occurrences). The largest effect is Role, with Card Holders producing approximately three times as many gestures as Non-Card Holders. The effect of Experience Type (i.e. target item) is illustrated in Table 2. Descriptions of Toothaches prompted the most gestures and Laughs the least.

The interaction between Experience Type and Role is illustrated in Table 5. The people describing an experience rely more on Abstract Descriptive than Iconic Gestures. The people listening to them, by contrast, show the reverse pattern, replying more on Iconic gestures than on Abstract Descriptives.

The second GEE analysis, reported in Table 4, shows the results for average gesture production time. This replicates the main effect of role with Cardholders producing longer, i.e. more sustained, gestures than Non Cardholders. The overall average duration of gestures for two different categories of gesture are not reliably different, nor are the four different Experience Types. However, the interaction between role and Experience Type is Replicated with Card Holders taking longer over Abstract Descriptives than Iconics and Non Cardholders showing the reverse pattern.

In order to test whether patterns of use of Iconic and Abstract Descriptive Gestures vary systematically with the different levels of the Experience Type variable i.e. from positive (1) to negative (4) four additional Univariate General Linear Model analyses were carried out on the frequency of gestures. It is important to note that the model fit is not as good for the current data and the statistical power is lower than the GEE analysis. However, this does provide a way to compare the trends for the different roles and different gesture types across the four levels of Experience Type.

For Cardholders, Iconic gestures show a reliable linear pattern of increase across from positive to negative experience ($F(3,55) = 3.47$, $p = 0.01$). However, Abstract Descriptive gestures do not show the same pattern ($F(3,55) = 1.01$, $p = 0.11$). For Non Cardholders neither Iconic gestures ($F(3,55) = 0.73$, $p = 0.28$) nor Abstract Descriptive Gestures ($F(3,55) = 0.06$, $p = 0.73$) show a reliable pattern of increase from positive to negative Experience types.

4 Discussion

Overall, there is little evidence of direct mimicry in this corpus, people listening to the description of an experience rarely produce gestures or expressions that match, in any simple way, those produced by the speaker. Patterns of listener gesture are systematically different from those of the people they are listening to.

The results show that speakers (Cardholder) perform more gestures, with a longer duration than listeners (Non Cardholder) for each item. This is expected, as the task structure ensures that the description of the Cardholders experience should take precedence, resulting in the speaker performing a more in depth gesticulation about the felt experience on the card to communicate the experience to the Non Cardholder.

The manipulation of (un)pleasantness of experience affects speakers and listeners differently. More Iconic gestures were performed than Abstract Descriptive gestures for both the Cardholder and the Non Cardholder. For example, when describing a toothache, the speaker might add iconic gestures describing the location of the pain by pointing to it, or perhaps detailing that they had to eat on the other side of their mouth, accompanied gesturally by performing a chewing motion and pointing to that side simultaneously.

When comparing the ratio of Iconic to Abstract Descriptive gestures, the Cardholder produced a higher ratio of Abstract Descriptive gestures than the Non Cardholder. We speculate that the higher amount of Iconic gesture suggests that both participants focus more on the concrete context of the situation surrounding the experience than the felt experience itself. Interestingly, the Non Cardholder's focus' even less on the sensation of the felt experience than the contextual aspects of the experience so would be more likely to mimic the contextual descriptions back to the Cardholder. We might speculate that the Non Cardholder actually avoids engaging with the description of the felt experience as the abstract descriptive gestures are not seen to mimic the Cardholders.

Returning to Chartrand and Barghs' (1999) work, their findings suggest that the mimicry of postures and gestures are a continual source of informa-

Note: CH = Cardholder, NCH = Non-Cardholder.

				95% Wald Confidence Interval	
Role	Gesture Type	Mean	Std. Error	Lower	Upper
CH	Abstract Descriptive	3.212131	.4086602	2.411172	4.013090
	Iconic	2.888671	.1616220	2.571898	3.205444
NCH	Abstract Descriptive	.629115	.1798742	.276568	.981662
	Iconic	1.145358	.2085097	.736687	1.554030

Table 5: Illustrating the Interaction Between Gesture Type and Role

tion throughout a social interaction, communicating messages indicating understanding and attention. The consensus appears to be coordination behaviour is related to empathy, rapport and liking, although some see mimicry as the cause and effect of empathic understanding. Chartrand and Bargh argue that individuals use behaviour mimicry as a communicative tool on a completely non-conscious level and that this mimicry usually leads to emotional convergence (Chartrand and Bargh, 1999). However, non-concious mimicry is not suggested in our results, as listeners did not tend to mimic the abstract descriptive gestures, if the mimicry was truly non-concsious in the way the perception-action link and our first hypothesis suggest, the listener would tend to mimic all gesture types as performed by the speaker, and this was not the case in this situation.

For the experiences with higher ranked experiences of (un)pleasantness, the amount and duration of gestures both increase for the Cardholder than the lower ranked experiences. However, the Non Cardholder was not affected, or showed no difference in terms of gestures, by the difference in rank. These results are consistent with a situation in which the person describing an experience will add iconic information to convey more negative experiences but do not add information about the felt experience. Listeners, by contrast, do not appear respond differently, in terms of gesture, to different degrees of (un)pleasantness. These results are also contrary to our second and third hypothesis, that the listener would produce more mimicry either for the more intense experiences, or for the more unpleasant or negative experiences, however as the listeners' responses were not affected in terms of rank at all, this is not consistent with either hypothesis. Also, this provides further evidence inconsistent with the perception-behaviour link, as again nonconscious mimicry would be affected by an increased frequency of gestures, as is the case with speakers' gestures in the higher ranked experiences, however listeners' gestures are not affected.

Why do listeners appear to avoiding mimicking the abstract descriptive gestures produced by the speaker? One alternative explanation is that this might be an issue of politeness, perhaps an acknowledgement of such inner experience is seen to be intrusive by the listener. This would result in the listener avoiding mimicking this description. A second alternative is that the interlocutors may find it easier to demonstrate the cause of the sensation or act out the outcome for the listener to understand the experience to infer or simulate how it felt, rather than describe the sensations itself.

An important limitation of the current study is that it looks at a limited number of gesture categories and doesn't examine their specific content or the structure of the interactional sequences in which they occur. Further work will examine the interactional context more closely and the other ways a listener can acknowledge or engage with speaker's descriptions of felt experience. For example, Bavelas et al. (1986) classify empathetic listener responses as motor mimicry. Here motor mimicry is not a straight mirroring or general imitation in the sense we have used the term mimicry in this paper, but is defined as the mimicry of an expressive behaviour, or the performance of the expected expressive behaviour of an occurrence in the perspective of another. Motor mimicry is found within a micro-social interaction where there is a high level of reciprocity and mutual influence between speaker and listener. Conceptualised as primitive empathy, motor mimicry is described as an automatic reflex of conditioned cues based on ones own prior experience. Bavelas and her colleagues suggest that motor mimicry serves as an expression of the perceived emotion, an interpersonal act to put across, in their words, I feel as you do (Bavelas et al., 1986). This is a response that acknowledges and engages with the felt experience, while not necessarily mimicking the Abstract Descriptive gesture that accompany the description as shown by this study, indicating a possible avenue for further work.

Acknowledgments

This research is supported by the Media and Arts Technology programme, an RCUK Doctoral Training Centre.

References

Bavelas, Janet. and Gerwing, Jennifer 2011. The Listener as Addressee in Face-to-Face Dialogue *International Journal of Listening* 25(3), 178-198.

Bavelas, Janet., Chovil, Nicole. and Lawrie, Douglas. 1992. Interactive gestures *Discourse Processes* 15, 469-489.

Bavelas, Janet., Black, Alex., Lemery, Charles., and Mullett, Jennifer. 1986. i show how you feel: Motor mimicry as a communicative act. *Journal of Personality and Social Psychology* 50(2):322329.

Beattie, Geoffrey and Shovelton, Heather. 2002. What properties of talk are associated with the generation of spontaneous iconic hand gestures? *The British journal of social psychology*, 41(3): 403-17.

Chartrand, Tanya. and Bargh, John. 1999. The chameleon effect: The perception-behaviour link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893910.

Clark, Herbert. 1996. Using Language. Cambridge: University Press.

Efron, David. 1941. *Gesture and environment.* Oxford: King's Crown Press.

Ekman, Paul. and Friesen, Wallace. 1969. The repertoire of nonverbal behaviour: Categories, origins, usage, and coding. *Semiotica* 1, 49- 98.

Gullberg, Marianne. 2009. Gestures in spatial descriptions. *Lund Working Papers in Linguistics* 47, 87-97.

Hatfield, Elaine., Cacioppo, John., and Rapson, Richard. 1993. Emotional contagion. *Current Directions in Psychological Sciences* 2, 96-99.

Hyden, Lars-Christer. and Peolsson, Michael. 2002. Pain Gestures: The Orchestration of Speech and Body Gestures. *Health* 6(3): 325-345

Inoue, Masashi., Ogihara, Mitsunori., Hanada, Ryoko. and Furuyama, Nobuhiro. 2010. Utility of Gestural Cues in Indexing Semantic Miscommunication. *Proceedings of the 5th International Conference on Future Information Technology*

Kendon, Adam. 2004. *Gesture: Visible action as utterance.* Cambridge: University Press.

Krauss, Robert., Chen, Yihsiu. and Chawla, Purnima. 1996. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? Mark. Zanna (Ed.), *Advances in experimental social psychology* San Diego, CA: Academic Press.

Marcus, Gregory., Cohen, Joshua., Varosy, Paul., Vessey, Joshua., Rose, Emily., Massie, Barry., Chatterjee, Kanu. and Waters, David. 2007. The utility of gestures in patients with chest discomfort. *The American journal of medicine* 120(1), 83-9

McNeill, David., Cassell, Justine and Levy, Elena. 1998. Abstract deixis. *Semiotica* 95(1/2), 5-19

McNeill, David. 1996. *Hand and mind: What gestures reveal about thought.* Chicago: University Press

Mol, Lisette., Krahmer, Emiel., Maes, Alfons. and Swerts, Marc. 2012. Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language* 66(1), 249-264

Nathan, Mitchell. and Havas, David. 2006. Intersubjectivity as a Basis for Gesture Production *Proceedings of the 28th Annual Conference of the Cognitive Science Society* 1821. Vancouver, Canada: Cognitive Science Society.

Parrill, Fey. and Sweetser, Eve. 2004. What we mean by meaning: Conceptual integration in gesture analysis and transcription. *Gesture*: 4(2), 197-219

Preston, Stephanie. and de Waal, Frans. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*: 25, 1-72.

Rowbotham, Samantha., Holler, Judith., Lloyd, Donna. and Wearden, Alison. 2011. How Do We Communicate About Pain? A Systematic Analysis of the Semantic Contribution of Co-speech Gestures in Pain-focused Conversations. *Journal of Nonverbal Behavior*: 36(1), 1-21

Schegloff, Emanuel. 1992. Repair after Next Turn : The Last Structurally Provided Defense of in Conversation Intersubjectivity *American Journal of Sociology*: 97(5), 1295-1345.

Wittgenstein, Lugwig. 1958. Philosophical Investigations: Translated by G.E.M. Anscombe. Oxford: Basil Blackwell

Dialogue Acts Annotation Scheme within Arabic discussions

Samira Ben Dbabis
ANLP Research Group
MIRACL Laboratory
University of Sfax,
TUNISIA
samira.benedbabis
@fsegs.rnu.tn

Fatma Mallek
ANLP Research Group
MIRACL Laboratory
University of Sfax,
TUNISIA
fsegs.fatmamallek
@gmail.com

Hatem Ghorbel
University of Applied
Science of West
Switzerland
HE-Arc Ingénierie
hatem.ghorbel
@he-arc.ch

Lamia Belguith
ANLP Research Group
MIRACL Laboratory
University of Sfax,
TUNISIA
l.belguith
@fsegs.rnu.tn

Abstract

Building discourse structure in human discussions needs a task of dialogue act annotation. In this paper, we propose dialogue act taxonomy in Arabic language. The proposed scheme is based mainly on the argumentative function that occurs frequently in debate conversations expressing opinions, ideas and arguments. To validate the reliability of our classification, we measure the agreement between two human annotators. Results show an average kappa score of 0.84 which expresses high reliability. To automatically generate annotated corpora, we developed an annotation tool that supports our dialogue act taxonomy for Arabic language.

1. Introduction

Dialogue Act (DA) annotation is a hot research topic in both human-to-human and human–computer speech communication. This task performed mainly in understanding the role of a user's utterance in the dialogue (Sridhar et al., 2009).

This field attracted researchers in linguistics (Austin, 1962; Searle, 1969) and computational linguistics (Core and Allen., 1997; Traum, 1999) since long time. Recent research on spoken dialogue processing has investigated computational dialogue act models of human-human and human-computer conversations (Stolcke et al., 2000).

The annotation task is fundamental to many studies in human discussions analysis as they reflect shallow discourse structures of language that can be investigated to build an argumentative structure of discussions.

Thus, the main goal of annotating DA in our work is to build adjacency pairs that reflect the DA sequences in Arabic discussions such as question/answer, opinion/reject, confirmation request/confirmation, etc.

These pairs are then investigated to generate an argumentative structure of the discourse that can help user to answer complex queries as "who rejected the proposal of M. X?".

The argumentative information level is based mainly on exchanging information, raising issues, expressing opinions, making suggestions, providing arguments, negotiating alternatives, and making decisions.

Thus tracking argumentative information is of central importance for building memories, browsing and summarizing discussions content.

To facilitate extracting argumentative data, it is useful to automatically annotate participant interaction characteristics specifically by identifying agreement and disagreement in order to understand social dynamics.

Annotating debate programs acts can be also a motivating task when a user needs information about a past discussion that he missed, or wants to recollect discussion dynamics (topic discussed, agreements, disagreements, arguments, etc).

In this perspective, we propose a dialogue act taxonomy including mainly argumentative actions related to acceptations, rejects, etc, in Arabic debate programs.

This paper is structured in four sections. First, we focus on the role of DA in building conversation structures. The next section exposes previous works in Dialogue Acts annotation research field and summarizes the main annotations schemes. In the third section, we experiment, with an empirical study, Arabic discussions and we propose our own DA taxonomy. Finally, we illustrate our annotation scheme by developing an annotating tool that generates annotated structures which can be used later as the basis of machine learning algorithms.

2. Argumentative Discourse Structure

Dialogue acts play a vital role in the identification of discourse structure.
In this context, (Grosz and Sidner, 1986) claim about task structure influencing dialogue structure. It seems likely that there are structures higher than a single utterance, yet more fine grained than a complete dialogue. Several researchers identify structures within dialogue at levels higher than individual utterances or speaker turns, but below the level of complete discourse description. There has been some significant exploration of the use of sequences of Dialogue Acts, at a number of levels of granularity.
The simplest dialogue sequence model is the use of adjacency pairs (Schegloff et al., 1973) which are functional links between pairs of utterances such as question/answer, opinion request/opinion, etc.
Within the adjacency pairs model, the importance of tracking a deeper structured representation based on argumentation theory has been recognized in (Pallota et al., 2004; Galley et al., 2004; Hillard, et al., 2003). These models help in constructing the argumentative information needed to express participants' intentions and to answer real user queries.

A simple but expressive model of an argumentative structure is the "Issue Based Information Systems" (IBIS) model, proposed by (Kunz and Rittel, 1970) and adopted as a foundational theory in some computer-supported collaborative argumentation systems. Thus, this model captures and highlights the essential lines of a discussion in terms of what issues have been discussed and what alternatives have been proposed and accepted by the participants.
In our context, the argumentative structure of discussions can be helpful in browsing topics discussed, decisions made, agreements and disagreements between participants.

3. DA Annotation Schemes overview

In order to standardize annotation tags, a proliferation of labelling schemes has been developed, often started from the topology suggested by Searle (Searle, 1969).
The granularity of DA annotation labels varies considerably from domain-specific to open-domain annotation task.

The MapTask project (Anderson et al., 1991), outlining task-oriented dialogues, is a collection of human conversations in which two people negotiate an agreed route on separate maps.
The MapTask labeling scheme uses 12 DA labels divided into two main categories: initiating and response moves.

Later, the Verbmobil project developed in Germany (1993-2000) aimed at the construction of an automatic speech to speech translation system for the languages German, American English and Japanese (Wahlster, 2000). A set of 43 DA is generated in a first phase (Jekat et al., 1995). These acts were organised in a hierarchy. There was a second phase of the Verbmobil project (Alexandersson et al., 1998), which expanded the dialogues from meeting scheduling to comprehensive travel planning. This domain change results a new hierarchy cluster of 18 top-level DA.

These schemas were all designed for specific-purpose application domains. They contained overlapping sets of communicative functions and made use of often mutually inconsistent terminology.

In the 1990s, a general-purpose schema called DAMSL: Dialogue Act Markup using Several Layers (Allen and Core., 1997; Core et al., 1998) is developed for multidimensional dialogue act annotation. With its focus on multidimensionality and domain-independence, this represented an important step forward in dialogue act annotation.
This annotation scheme leads to considering specific dimensions such as: communicative status, information level, forward-looking function and backward-looking function.
Several extensions of the DAMSL schema have been constructed for specific purposes, such as Switchboard-DAMSL (Jurafsky et al., 1997).

The comprehensive DIT++ schema (Bunt, 2006; Bunt, 2009) combines the

multidimensional DIT schema, developed earlier (Bunt, 1994) with concepts from these various alternative schemas, and provides precise and mutually consistent definitions for its communicative functions and dimensions.
There are 11 dimensions of the DIT++ tag-set, with around 95 communicative functions, around 42 of which, like switchboard are for general purpose functions, whereas others cover elements of feedback, interaction management and the control of social obligations.

These annotation schemes have been used to mark-up several dialogue corpora in different languages such as English, German and Spanish.
However, very few works were developed for Arabic. To our knowledge, there is only one work achieved at Memphis University (Shala et al., 2010) that proposes speech acts classification model including the following set of predefined categories: *assertion, declaration, denial, expressive evaluation, greeting, indirect request, question, promise/denial, response to question*, and *short response*.
This tag set includes general-purpose actions that can be applied to independent domain corpora.
Nevertheless, these acts are incomplete to build discourse structure and are unable to describe argumentative structure.
In fact, this taxonomy cannot annotate argumentative actions related to exchanging opinions, arguments, acceptations, rejects, etc.

4. Empirical Study

4.1. Experimental data

The corpus used to perform the experiments is a set of transcriptions of debate programs taken from "AL JAZEERA"[1] Arabic channel. It consists of human-human discussions about generic topics. The choice of this corpus is argued by the strong and intense argumentation hold in its content mainly conveyed by exchanging opinions, acceptations, rejects, etc.
The study corpus has been manually annotated at the dialogue act level by two human experts. Each discussion turn was manually segmented into utterances. Each discussion contains about 400 utterances with an average duration of 2 hours. Each utterance was assigned one label.

[1] www.aljazeera.net

4.2. Arabic taxonomy

In Arabic language, semantics "علم المعاني" include *statement* *"الخبر"* and *construction* "*الإنشاء*"

- *Statement الخبر*

In general, a sentence or phrase that is a statement can be said to be true or false.
A statement makes a claim about the world, and tries to change the belief of the listener. It generally refers to assertions, declarations following the representative class of Searle's taxonomy.

- *Construction الإنشاء*

Opposed to statement, construction includes actions that do not support to be true or false.
Two main categories are defined under the constructions: "الإنشاء الطّلبي" and "الإنشاء الغير طّلبي" referring respectively to *request* and *non request construction*.
Request construction can be expressed into questions, orders, etc, whereas non request category refers generally to exclamation, praising or complaint.
Actions included into these two subcategories are summarized in table 1.

Request Construction	**الإنشاء الطلبي**
Question	الإستفهام
Call	النداء
Polite Request	الإلتماس
Incitation	التحضيض
Order	الأمر
Discourage	النهي
Promise	الوعد
Hope	الترجّي
Wish	التمني
Invocation	الدعاء
Warning	التحذير
Non Request Construction	**الإنشاء الغير طلبي**
Exclamation	التّعجب
Praise	المدح
Complaint	الذّم

Table 1. Construction taxonomy

4.3. Our Dialogue Act taxonomy

Starting from tags used in previous annotation schemes, we propose a dialogue acts taxonomy that enables the labelling of argumentative acts that are fundamental in generating argumentative structure of human conversations. Thus, our empirical analysis leads to identifying five main groups of dialogue acts: *Social Obligation Management, Turn Management, Argumentative, Request, Answer* and *Statement*. We eliminate the non request category cause of its very few occurrence in the studied corpus.

The given categories can be applied for other languages and can be common across annotation schemes especially those tracking argumentative data.

Our taxonomy, following the same partition, is detailed in the next sections.

- *Social Obligation Management*

This category includes conventional acts such as opening, closing and greetings, in addition to the expressive acts following Searle's classification as thanking, apology, regret, etc.

- *Turn Management*

Turn management acts are used to elicit and provide feedback in order to perform turn speaking management in the discussion.

- *Request*

This category includes different request categories (confirmation request, explanation request, etc) and takes different forms (question, order, hope, wish, etc). This class includes initiatives often called forward-looking acts.
Request utterances can express several kinds of demands such as confirmation request, explanation request, justification request and opinion request. These tags are generally associated respectively to the following acts: confirmation, explanation, justification and opinion acts.

- *Argumentation*

Argumentation is mainly based on exchanging opinions, accepting or rejecting others ideas. It's the fact to convince others by giving arguments, explanations, examples …
Thus, *argumentative acts* represent the core acts in the discussion that express argumentative actions.

- *Answer*

Answers consist of general-purpose acts that reply to questions. This category often represents the backward-looking function. It is generally paired with the question label.

- *Statement*

Statement label describes non opinion statement that can state an event or an assertion.

- *Other*

includes non-interpretable and non-classifiable utterances.

4.4. Kappa Ratio

A first step in determining the quality of a set of annotations is to evaluate the agreement between annotators.
The current standard metric used for measuring inter-annotator agreement in classification tasks is the Cohen kappa statistic (Carletta, 1996). This metric can be used effectively only on break classifications when the number of segments is unconstrained. Also, this metric does not adequately accommodate near-miss topic break assignments and other desired tolerances for slightly differing segmentations. In this section, we present the results of evaluating inter-annotators agreement.
First, we take discussions that have been segmented identically. Then, we appoint two human experts to annotate separately the tokenized conversations while following our classification taxonomy guidelines.
The agreement between annotators is calculated using the kappa measure. We obtain an average score of 0.84. This inter-annotator agreement ratio expresses high reliability between human annotators. The main inter-class differences are between Argumentative and Turn Management labels. For instance the word *"نعم"* (*yes*) in the example below can express an acceptation, a confirmation or just a backchannel action to continue the discussion.

Annotation 1:

Act1: لا نعتقد بأن تونس مهيأة لنظام عسكري أصلا

We don't believe that Tunisia is really ready to a military regime.

<Class="Argumentation", DA="opinion">

Act2: نعم

yes.

<Class="Argumentation", DA="Acceptation">

Annotation 2:

Act1: لا نعتقد بأن تونس مهيأة لنظام عسكري أصلا

Social Obligation Management		
Opening	اداب الافتتاح	Dialogue beginning
Closing	اداب الإقفال	Dialogue ending
Greeting	تحية	Addressee's salutation
Polite Formula	إلتماس	Showing regard for others, in manners, speech, behaviour, etc.
Introduce	تقديم	Self-introduction, speakers and topics introduction.
Thanking	شكر	Gratitude feeling
Apology	اعتذار	Regret having made an error in understanding, evaluating or executing an utterance
Regret	تأسف	Feeling of sorry and disappointment.
Turn Management		
Acknowledgement	قبول	Agreement with previous utterance or addressee understanding feedback.
Calm	تهدئة	Calming down speakers to control the situation.
Clarify Request	طلب توضيح	Asking addressee for reformulation/repetition of previous utterance for clarification.
Clarify	توضيح	Reply to a clarification request
Feedback	تذكير	Remind addressees about what was evoked in previous utterances
Out of topic	أسلوب الحكيم	A way to change the topic and to bypass the addressee's question
Non understanding signal	علامة عدم فهم	Expressing non understanding of the previous utterance
Request		
Question	استفهام	Asking for information from the addressee and requiring a reply.
Order	أمر	Direct request obliging addressee to do something
Promise	وعد	Potentially promising for achieving a certain goal
Hope	ترجّي	feeling that something desired may happen
Wish	تمني	Longing for something with expectation of its fulfillment
Invocation	دعاء	Prayer that implies to call upon God
Warning	تحذير	Desisting from a specified undesirable action
Argumentation		
Opinion	إبداء الرّأي	Subjective belief that may be supported by arguments
Appreciation	استحسان	Favorable judgment or opinion
Disapproval	استهجان	Feeling of disliking something or what someone is doing
Accept	موافقة	Affirmative answer expressing agreement with addressee
Partial Accept	موافقة جزئية	Expressing partial agreement with addressee
Reject	رفض	Refusal to accept addressee's opinion, judgment or proposal
Partial Reject	رفض جزئي	Partial disagreement with addressee opinion, judgment or proposal
Argument	حجّة	Attempt to persuade someone of something, by giving reasons or evidence for accepting a particular conclusion.
Justification	تعليل	Defending by reasoning an action or a belief
Explanation	تفسير	Making something comprehensible by describing the relevant structure or operation or circumstances
Confirmation	توكيد	Additional proof that something which was believed is correct
Conclusion	استنتاج	Decision or opinion or judgment reached after consideration
Answer		Reply to a question
Statement		Affirming or asserting or stating something

Table 3. Our Dialogue Act Taxonomy

We don't believe that Tunisia is really ready to a military regime.

<Class="Argumentation", DA="opinion">

Act2: نعم

yes.

<Class="Turn Management", DA="Acknowledgement">

In order to detail the intra-class reliability, we calculate the kappa score within each class (see Table 2).

Class	Kappa
Request	0.96
Social Obligation Management	0.90
Turn Management	0.82
Argumentative	0.66

Table 2. Kappa per class

Most categories seem to be labelled fairly reliably such as *Social Obligation Management* reaching a kappa ratio of 0.9 and *Turn Management* category with kappa = 0.82.

However, experienced annotators scored a kappa of 0.66 for the *Argumentative* dimension. This rate decline can be justified by the difficulty of annotating argumentative tags. Major experts' annotation differences are focused on ambiguities in labelling opinion tags, in detecting arguments.
Besides, an utterance can refer to more than one action such as <argument, explanation>, <opinion, argument>. In these cases, human annotations could be different given that experts should assign only one label to each utterance.
Annotators' disagreement when annotating argumentative dialogue acts can be explained by the lack of linguistic markers. For instance, "explanation", and "justification" acts can be ambiguous for annotation especially when they are used without specific cue words such as "يعني" (*"that means"*), generally used for explanation, and "لأن" ("*because*") often followed by a justification.
Moreover, the word "يعني" (*"that means"*) used mostly as an explanation cue word can be a trivial expression often used in spoken dialogue as detailed in the following example.

Annotation 1:

يعني جميل أن يكون الإنسان منحازا للشعب

That means it is great when a human is aligned with people.

< Class="Argumentation", DA="Explanation">

Annotation 2:

يعني جميل أن يكون الإنسان منحازا للشعب

That means it is great when a human is aligned with people.

< Class="Argumentation", DA="Appreciation">

In order to construct a training corpus for machine learning classification, we intend to reach a minimum of kappa score of 0.6.

5. ActAAr Annotation Tool

Dialogue acts annotation task requires a considerable effort from human annotators. Therefore, many annotating tools have been developed to offer more interaction with annotated corpora.
In fact there are numerous tools for general annotation tasks such as GATE and MATE and other tools for dialogue act annotation like XDML and DAT.
GATE system (Cunningham et al., 2002) is one of the most commonly used systems. It supports manual annotation, information extraction, semi-automatic semantic annotation, etc.
MATE [2]workbench (Klein, 1999) is a multimodal annotation tool. It can be used with different annotation schemes in XML format. It also allows the corpus designer to write rule based transformations using a language very similar to XSLT.
XDML (eXtensible Dialogue Markup Language Tool) was designed for annotating transcribed dialogues according to semantic, functional and stylistic characteristics. It was developed within the AMITIES[3] project.
DAT is a Perl/Tk tool for dialogue act tagging which processes files in SGML format. It was developed in the DAMSL[4] project. It directly supports dialogue structures (turns and utterances) and includes data from different modalities.

As presented above, these tools are not suitable for dialogue act annotation in Arabic language. Besides, DAT and XDML tools were

2 http://mate.nis.sdu.dk
3 http://www.dcs.shef.ac.uk/nlp/amities/amitiesdemos.htm
4 http://www.cs.rochester.edu/research/speech/damsl/

developed for specific purposes within annotation projects.

GATE and MATE tools are not simple for use by human annotators as they need more proficiency and effort to be used to their annotation guidelines.

Therefore, we have developed an annotation tool named ActAAr (**Act**s **A**nnotation in **Ar**abic) which is simple to use and supports our dialogue act taxonomy.

In fact, our tool is a java application for dialogue acts annotation in Arabic discussions. It uses the taxonomy detailed in table 3.
Indeed, the expected input format is plain text discussions files. The loaded file is then automatically segmented into turns. After the user's annotation, the output structure is saved in an XML labeled file tokenized into functional units (turns and utterances). For each utterance, the output tag includes the DA's label and class.
The annotation process is done by the two following tasks:

1- Select the utterance: the user selects the text by using the mouse from the dialogue shown in the left side of the screen.
2- Select the relevant dialogue act: the user chooses the appropriate class from the list shown. Then he selects one dialogue act from the selected class by a simple mouse-right-click.

When these two tasks are carried out, the program adds the following tag: *<utterance ID="n1" DA="d1" Class="c1"* > under the corresponding turn (see figure 1).

<Turn ID ="79" Speaker="غسان بن جدو">
<Utterance ID="50" DA=" Question" Class=" Request ">
أين؟
</utterance>
<Turn ID ="80" Speaker="راشد الغنوشي">
<Utterance ID="51" DA=" Answer" Class=" Answer ">
موجودة في الدستور
</Utterance>
</Turn>
<Turn ID ="81" Speaker="غسان بن جدو ">
<Utterance ID="52" DA="Acknowledgement" Class =" Turn_Management">
طيب
</Utterance>
<Utterance ID="53" DA="Polite_formula" Class=" Social_Obligation_Management">
عفوا شيخ راشد
</Utterance>
<Utterance ID="54" DA=" Explanation_Request" Class=" Request">
يعني هل أفهم من ذلك الآن عندما تقول بهذه الطريقة إنك الآ ن
تدعو إلى تعديل الدستور أو إلغاء الدستور أو ماذا؟
</Utterance>
</Turn>
<Turn ID ="82" Speaker="غسان بن جدو ">
<Utterance ID="55" DA="Explanation" Class =" Argumentative">
أنا أدعو لأكثر من ذلك أدعو إلى تفكيك منظومة الاستبداد
</Utterance>
<Utterance ID="56" DA="Conclusion" Class =" Argumentative">
ولذلك أنا أدعو أهل الانتفاضة أدعو شعب تونس إلى تسلم اليقظة
وتسلم الحضور
</Utterance>
</Turn>

Figure 1. An annotated corpus sample

6. Conclusion and future work

In this paper we have proposed a Dialogue Acts scheme for argumentative annotation of Arabic discussions. We evaluated the reliability of this scheme by manually annotating a corpus of debate programs transcriptions and assessing the inter-annotator agreement using the Kappa measure. From the obtained results, we can conclude that the proposed taxonomy is fairly reliable and at the current stage needs to be refined in order to obtain better agreement. However, we noticed that some disagreement might be due to our under-constrained guidelines that do not provide clear criteria for discriminating between possible categories.
As a future work, we intend to improve the annotation guidelines by providing a set of mark-up labels and the rules for their application. These guidelines will be the basic reference for human annotators to generate coherent annotations of discussions.
From a practical point of view, we intend to integrate navigation and research modules that extract statistics from annotated corpora (DA frequency in the corpus, acts per class, adjacency pairs, etc.).
Finally, we will use our annotation tool to generate a large number of annotated structures which can be used later as a basis of a machine learning algorithm in automatic annotation task.

References

(Austin, 1962) Austin, J. L. 1962. How to Do Things with Words. Oxford University Press, Oxford.

(Alexandersson et al., 1998) Alexandersson, J., B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. 1998. Dialogue Acts in VERBMOBIL-2 (second edition). Vm report 226, DFKI GmbH, Universities of Berlin, Saarbrcken and Stuttgart.

(Allen and Core, 1997) Allen, J. and M. Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. Technical report, Jan.

(Anderson et al., 1991) Anderson, A., M. Bader, E. Bard, E. Boyle, and G. Doherty. 1991. The HCRC Map Task Corpus. Language and Speech, 34:351-366.

(Bunt, 1994) Bunt, H. 1994. Context and Dialogue Control. THINK, 3:19-31.

(Bunt, 2006) Harry Bunt. 2006. Dimensions in dialogue annotation. In Proceedings of LREC 2006.

(Bunt, 2009) Harry Bunt. 2009. 'The DIT++ taxonomy for functional dialogue markup'. In Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009), Dirk Heylen, Catherine Pelachaud, Roberta Catizone, and David Traum, editors, Budapest, May 12, 2009.

(Carletta et al., 1996) Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2), 249-254.

(Core and Allen, 1997) Core, M. and J. Allen. 1997. Coding Dialogs with the DAMSL annotation scheme. In AAAI Fall Symposium on Communicative Action in Humans and Machines, MIT, Cambridge, MA.

(Core et al., 1998) Core, M., M. Ishizaki, J. Moore, C. Nakatani, N. Reithinger, D. Traum & S. Tutiya (1998) Report of The Third Workshop of the Discourse Resource Initiative, May 18-22, 1998, Chiba University, Chiba, Japan.

(Cunningham et al., 2002) Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE:A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.

(Galley et al., 2004) Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *ACL 2004, Barcelona*, 2004.

(Grosz et Sidner, 1986) Grosz, B. and C. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. Computational Linguistics, 19(3).

(Hillard, et al., 2003) D. Hillard, M. Ostendorf, and E. Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In Proceedings of HLT/NAACL.

(Jekat et al., 1995) Jekat, S., R. Klein, E. Maier, I. Maleck, M. Mast, T. Berlin, and J. J. Quantz. 1995. Dialogue Acts in VERBMOBIL. Vm report 65, DFKI GmbH, Universities of Berlin, Saarbrcken and Stuttgart.

(Jurafsky et al., 1997) Jurafsky, D., R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. 1997. Automatic Detection of Discourse Structure for Speech Recognition and Understanding. In Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara.

(Klein, 1999) Klein, M. 1999. Standardisation E_orts on the Level of Dialogue Act in the MATE Project. Proceedings of the ACL Workshop Towards Standards and Tools for Discourse Tagging.

(Kunz and Rittel, 1970) Kunz W. and Rittel H. W. J. (1970). *Issues as elements of information systems*. Technical Report WP-131, Berkeley: University of California.

(Pallota et al., 2004) Pallotta, V., Ghorbel, H., Ruch, P., and Coray, G. (2004). An argumentative annotation schema for meeting discussions. In Proceedings of the 4th International Conference on Language Resources (LREC 2004), May 26-28, 2004, Lisbon, Portugal, pages 1003-1006.

(Schegloff et al., 1973) Schegloff, E. A. and H. Sacks. 1973. Opening Up Closings. Semiotica, 7:289-327.

(Searle, 1969) Searle, J. R. 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

(Shala et al., 2010) Shala, L., Rus, V., & Graesser, A. C. (2010). Automated speech act classification in Arabic. Subjetividad y Procesos Cognitivos, 14, 284-292.

(Sridhar et al., 2009) V.K.R Sridar, S. Bangalore, and S.S. Narayanan. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. Computer Speech & Language, 23(4): 407-422. Elsevier Ltd.

(Stolcke et al., 2000) Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of

Conversational Speech. In Computational Linguistics 26(3), 339-373.

(Traum, 1999) Traum, D. (1999) Speech Acts for Dialogue Agents. In: M. Wooldridge & A. Rao (eds.) Foundations of rational agency, Kluwer, Dordrecht, 169–201.

(Wahlster, 2000) Wahlster, W. 2000. Verbmobil: Foundations of Speech-To-Speech Translation. Springer.

Declarative Design of Spoken Dialogue Systems with Probabilistic Rules

Pierre Lison
Department of Informatics
University of Oslo, Norway
plison@ifi.uio.no

Abstract

Spoken dialogue systems are instantiated in complex architectures comprising multiple interconnected components. These architectures often take the form of pipelines whose components are essentially black-boxes developed and optimised separately, using ad-hoc specification formats for their inputs and outputs, domain models and parameters.

We present in this paper an alternative modelling approach, in which the dialogue processing steps (from understanding to management and to generation) are all declaratively specified using the same underlying formalism. The formalism is based on *probabilistic rules* operating on a shared belief state. These rules are expressed as structured mapping between state variables and provide a compact, probabilistic encoding for the dialogue processing models. We argue that this declarative approach yields several advantages in terms of transparency, domain-portability and adaptivity over traditional black-box architectures. We also describe the implementation and validation of this approach in an integrated architecture for human-robot interaction.

1 Introduction

Spoken dialogue systems typically rely on complex pipeline architectures, including components such as speech recognition, semantic parsing, dialogue act classification, dialogue management, sentence planning, surface realisation and speech synthesis, in addition to extra-linguistic modules for e.g. situation awareness or the execution of physical actions, see e.g. (Bos et al., 2003; Bohus et al., 2007; Kruijff et al., 2007). In many cases, these components are developed separately and rely on dedicated representation formats for their inputs and output variables, internal models and parameters.

For instance, a dialogue act classifier might take as input a N-Best list of recognition hypotheses, and outputs the corresponding dialogue act(s), using a set of shallow patterns as internal model to relate the input to the output. Similarly, a dialogue manager takes a given dialogue state as input, and outputs the optimal action (if any) to perform in such state based on a specific planning procedure.

These representation formats are unfortunately rarely compatible with one another, which makes it difficult to derive a semantic interpretation for the dialogue state as a whole (in terms e.g. of a joint probability distribution). Moreover, much of the task knowledge is typically encoded in procedural form within the component workflow, reducing the system portability to other domains due to the necessity of reprogramming some of the modules.

We present in this paper an alternative approach to the specification and optimisation of the various models used in a dialogue system architecture. The key idea is to declaratively specify the task-specific models using a shared, generic representation formalism, and strip down the system architecture to a small set of core algorithms for updating the dialogue state based on these models. The representation formalism we describe in this paper is based on the concept of a *probabilistic rule*. These rules are expressive enough to capture the structure for most processing tasks, from dialogue understanding to management and to generation. Moreover, they can be either manually designed or have their parameters estimated from data.

This declarative approach to the design of spoken dialogue systems has several advantages. The first one pertains to domain portability. Given that the dialogue architecture is essentially reduced to a generic platform for rule instantiation and inference, porting the system to a new domain only requires a rewrite or extension of the domain-specific rules, without having to reprogram a single component. It also provides a more transparent semantics for the system as a whole, since all state variables are described and related to each other in a unified and theoretically sound framework, grounded in probabilistic inference. Finally, the use of probabilistic rules enables the construction of very flexible processing pipelines, by allowing state variables to depend or influence each other in any order and direction. The system designer is thus free to combine in the same architecture both deep and shallow semantic parsers for dialogue understanding, or both reactive and deliberative policies for dialogue management.

The architecture revolves around a shared dialogue state, encoded as a Bayesian Network including all variables relevant for the interaction. The use of a Bayesian Network allows us to account for the various kinds of uncertainties arising in spoken dialogue (speech recognition errors, unknown user intentions, etc.) as well as the conditional dependencies between state variables. At runtime, this dialogue state is continuously updated via probabilistic rules. As we shall see, these rules are instantiated by extending the Bayesian network with new nodes and conditional dependencies.

We showed in our previous work how to estimate the parameters of these models given limited amounts of Wizard-of-Oz training data (Lison, 2012). The present paper builds upon this approach, but concentrates on the design and specification of these probabilistic rules for various processing tasks, leaving out the question of parameter estimation. Hence, we will simply assume through this paper that the parameters have been already assigned, either from training data or expert knowledge.

The rest of the paper is as follows. We first provide generalities about Bayesian Networks and dialogue models. We then describe our approach by defining the probabilistic rules and their use in the dialogue processing workflow. We also detail a system implemented for a human-robot interaction domain, which exploits probabilistic rules to perform tasks related to dialogue understanding, management and generation. Finally, we discuss and relate our approach to previous work, and conclude.

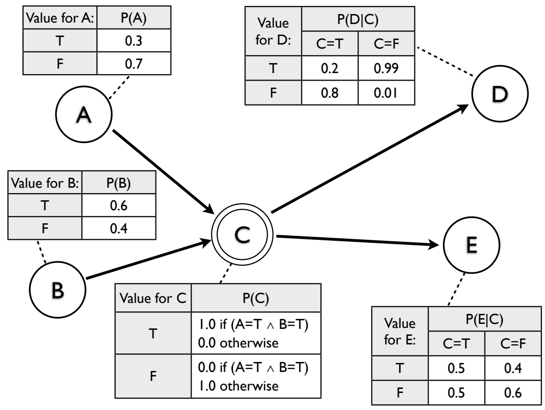


Figure 1: Bayesian network with 5 nodes. An example of query on this network is $P(A{=}\mathrm{T}|D{=}\mathrm{T}) \approx 0.18$.

2 Background

2.1 Bayesian Networks

The probabilistic models used in this paper are expressed as directed graphical models, also known as Bayesian Networks. Let $X_1...X_n$ denote a set of random variables. Each variable X_i is associated with a range of mutually exclusive values. In dialogue models, this range is often discrete and can be explicitly enumerated: $Val(X_i) = \{x_i^1, ..., x_i^m\}$.

A Bayesian Network defines the conditional dependencies between variables using a directed graph where each node corresponds to a variable X_i. Each edge $X_i \rightarrow X_j$ denotes a conditional dependence between the two nodes, in which case X_i is said to be a *parent* of X_j. A conditional probability distribution $P(X_i|Par(X_i))$ is associated with each node X_i, where $Par(X_i)$ denotes the parents of X_i.

Conditional probability distributions (CPDs) can be defined in various ways, from look-up tables to more advanced distributions (Koller and Friedman, 2009). Together with the directed graph, the CPDs determine the joint probability distribution $P(X_1...X_n)$. The network can then be used for inference by querying the distribution of a subset of variables, often given some additional evidence, as illustrated by the example in Figure 1.

2.2 Dialogue models

The dialogue state $\mathbf{s}$ can generally be decomposed into a set of state variables $\mathbf{s} = \{s_1, ...s_n\}$ (Williams and Young, 2007). Each state variable represents a relevant feature of the interaction and its context. For instance, the state variables for a human-robot interaction scenario might be composed of tasks to accomplish, the interaction history, past events, as well as objects, spatial locations and agents in the environment. A minimal dialogue state can be defined as $\mathbf{s} = \langle u_u, a_u, i_u, a_m, u_m, c \rangle$, where u_u is the last user utterance, a_u the last dialogue act, i_u the current user intention, a_m the last system act, u_m the last system utterance, and c the context.

Due to uncertainty, many variables are only partially observable. We thus encode our knowledge of the current dialogue state in a distribution $\mathbf{b}(\mathbf{s}) = P(s_1, ..., s_n)$ called the *belief state*, conveniently expressed as a Bayesian Network (Thomson and Young, 2010). This belief state $\mathbf{b}$ is regularly updated with new information. The workflow illustrated in Figure 2 can then be formalised in terms of inference steps over this belief state:

1. Upon detection of a new speech signal, the speech recogniser generates the N-best list of recognition hypotheses $\tilde{\mathbf{u}}_\mathbf{u} = P(u_u|o)$;
2. Speech understanding then estimates the most likely dialogue act(s) realised in the utterance: $\tilde{\mathbf{a}}_\mathbf{u} = P(a_u|\mathbf{b})$;
3. The user intention is updated with the new interpreted dialogue act: $\tilde{\mathbf{i}}_\mathbf{u} = P(i_u|\mathbf{b})$;
4. Based on the updated belief state, the action selection searches for the optimal system action to perform: $a^*_m = \arg\max_{a_m} Q(a_m|\mathbf{b})$;
5. The system action is then realised in an utterance u_m, which is again framed as a search for $u^*_m = \arg\max_{u_m} Q(u_m|\mathbf{b})$;

The models defined above use $P(x|\mathbf{b})$ as a notational convenience for $\sum_{\mathbf{s^i} \in Val(\mathbf{s})} P(x|\mathbf{s}{=}\mathbf{s^i})\mathbf{b}(\mathbf{s^i})$. The sequence above might be adapted in various ways depending on the domain. A basic reactive system might for instance ignore the user intention and directly select its actions based on the last dialogue act. Similarly, user-adaptivity might be captured via additional processing steps to estimate and exploit features related to the user model.

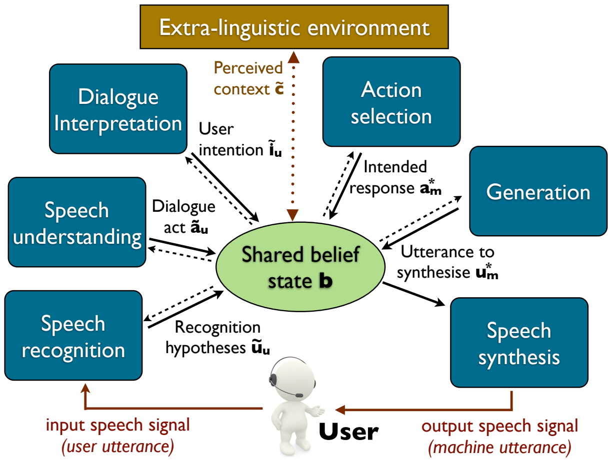


Figure 2: Dialogue system architecture schema.

3 Approach

The starting point of our approach is to express the probabilistic dialogue models described above using a compact encoding which takes advantage of the *internal structure* present in most processing tasks. This structure can take several forms:

- The probability (or utility) of a given output variable typically depends on only a small subset of input variables, although the number and identity of these variables might naturally differ. The state variable encoding the physical location of a mobile robot is for instance relevant for answering a user requesting its location, but not for responding to a greeting act.

- Moreover, the values for the dependent variables can often be grouped into a small number of *partitions* yielding similar outcomes, thereby reducing the dimensionality of the problem. The partitions can be expressed via logical conditions on the variable values.

Probabilistic rules provide a general framework for expressing this internal structure. The rules express the model distribution in terms of structured mappings between input and output variables. At runtime, these rules are then applied on the belief state, thereby extending the Bayesian Network with new nodes and conditional dependencies. This updated Bayesian Network can then be directly used for inference, e.g. to compute the marginal distribution of a particular variable or to search for the optimal action to perform. The probabilistic rules thus function as high-level *templates* for the incremental construction of a classical probabilistic model.

3.1 Definitions

Rules can be of two possible types:

1. *probability rules*, which define probability models of the form $P(\mathbf{X}|\mathbf{Y})$, where $\mathbf{X}$ and $\mathbf{Y}$ both denote arbitrary sets of state variables;

2. *utility rules*, defining utility models of the form $Q(\mathbf{X}, \mathbf{A})$, where $\mathbf{X}$ represent a set of state variables, and $\mathbf{A}$ a set of action variables. The rule defines here the utility of a particular action sequence in $\mathbf{A}$ given the state defined by $\mathbf{X}$.

A rule is essentially defined as a condition-effect mapping, where each condition is mapped to a set of alternative effects. Depending on the type of rule, each effect will be assigned to either a *probability* or a *utility* value. The list of conditions is ordered and takes the form of a "**if** ... **then** ... **else**" case expressing the probability/utility distribution of the output variables depending on the inputs.

Probability rule

Formally, a probability rule r is defined as an ordered list of cases, where each case is associated with a condition c_i as well as a distribution over stochastic effects $\{(e_i^1, p_i^1), ..., (e_i^k, p_i^k)\}$. For each stochastic effect e_i^j, we have that $p_i^j = P(e_i^j|c_i)$, where $p_i^{1...m}$ satisfy the usual probability axioms. The rule reads as such:

if (c_1) **then**

$$\{P(e_1^1) = p_1^1, \; ... \; P(e_1^k) = p_1^k\}$$

...

else if (c_n) **then**

$$\{P(e_n^1) = p_n^1, \; ... \; P(e_n^m) = p_n^m\}$$

A final **else** case is implicitly added to the bottom of the list, and holds if no other condition applies. If not overridden, the default effect associated to this last case is void – i.e. it causes no changes to the distribution over the output variables.

Utility rule

Utility rules are defined similarly. Each case specified in the rule is associated to a condition c_i and a utility distribution over possible action sequences $\{(a_i^1, q_i^1), ..., (a_i^k, q_i^k)\}$, where a_i^j is a value assignment for a set of action variables, and $q_i^j = Q(c_i, a_i^j)$. The rule reads as:

if (c_1) **then**

$$\{Q(a_1^1) = q_1^1, \; ... \; Q(a_1^k) = q_1^k\}$$

...

else if (c_n) **then**

$$\{Q(a_n^1) = q_n^1, \; ... \; Q(a_n^m) = q_n^m\}$$

The default utility value of an action is set to 0. When several rules define a utility value for the same action, these utilities are *summed*.

Conditions

For both rule types, the conditions are expressed as logical formulae grounded in the input variables. They can be arbitrarily complex formulae connected by conjunctive, disjunctive and negation operators. The conditions on the input variables can be seen as providing a compact partition of the state space to mitigate the dimensionality curse. Without this partitioning in alternative conditions, a rule ranging over m variables each of size n would need to enumerate n^m possible assignments. The partitioning with conditions reduces this number to p mutually exclusive partitions, where p is usually small.

A wide range of conditional tests can be devised. In our implementation, the rule conditions for speech understanding were for instance expressed in terms of regular expressions matches on the user utterance u_u. Generally speaking, a condition is simply defined as a function mapping state variable assignments to a boolean value.

Effects

The rule effects are defined similarly: given a condition holding on a set of input variables, the associated effects define specific *value assignments* for the output variables. The effects can be limited to a single variable or range over several output variables. The effect can also be void, i.e. trigger no change to the distribution over output values.

Each effect is assigned to a scalar value defining its probability or utility, and several alternative stochastic effects can be defined for the same case. If a unique effect is specified, it is then implicitly assumed to hold with probability 1. These values are parameters which can be either hand-coded or estimated from data.

Example

Assume an action selection model structured with probabilistic rules, which operates on a belief state $\mathbf{b}$ containing the last user act a_u as well as a collection of objects perceived in the environment. The response to a polar question such as "is the object red?" can be captured by the following rules:

$$
\begin{aligned}
r_1 : \ & \textbf{if } (a_u\!= \mathrm{VerifyColour(o, c)} \\
& \qquad \wedge\ \mathrm{o.colour}\!=\!\mathrm{c}) \ \textbf{then} \\
& \quad \{Q(a_m\!= \mathrm{Confirm}) = 5\} \\
& \textbf{else} \ \ \{Q(a_m\!= \mathrm{Confirm}) = -4\}
\end{aligned}
$$

$$
\begin{aligned}
r_2 : \ & \textbf{if } (a_u\!= \mathrm{VerifyColour(o, c)} \\
& \qquad \wedge\ \mathrm{o.colour}\!\neq\!\mathrm{c}) \ \textbf{then} \\
& \quad \{Q(a_m\!= \mathrm{Disconfirm}) = 5\} \\
& \textbf{else} \ \ \{Q(a_m\!= \mathrm{Disconfirm}) = -4\}
\end{aligned}
$$

$$
\begin{aligned}
r_3 : \ & \textbf{if } (a_u\!= \mathrm{VerifyColour(o, c)}) \ \textbf{then} \\
& \quad \{Q(a_m\!= \mathrm{SayDontKnow}) = 1\} \\
& \textbf{else} \ \ \{Q(a_m\!= \mathrm{SayDontKnow}) = -2\}
\end{aligned}
$$

The rule specifies the following behaviour: if there is a reasonable certainty that the object is (resp. is not) of the correct colour, the system should confirm (resp. disconfirm). In case of uncertainty, it should utter "I don't know". The trade-offs between these actions are encoded by the utility parameters.

To illustrate the rules, assume that the dialogue state contains the two independent variables a_u and $\mathrm{o_1.colour}$, with the respective distributions:

$$
\begin{aligned}
& P(a_u\!=\!\mathrm{VerifyColour(o_1, blue)}) = 0.8) \\
& P(a_u\!=\!\mathrm{VerifyColour(o_1, black)}) = 0.2) \\
& P(\mathrm{o_1.colour}\!=\!\mathrm{blue}) = 0.75) \\
& P(\mathrm{o_1.colour}\!=\!\mathrm{green}) = 0.25)
\end{aligned}
$$

It is then trivial to calculate that, in this setting, the best action to perform is $\mathrm{Confirm}$, which has a utility $Q = 1.4$, while $Q(\mathrm{SayDontKnow}) = 1$ and $Q(\mathrm{Disconfirm}) = -0.4$.

3.2 Processing workflow

To ease the design of the architecture, the probabilistic rules are grouped into *models*. A model consists of a set of rules and the specification of a "trigger" variable which causes the activation of the model. For instance, the trigger for the speech understanding model $P(a_u|u_u)$ is the user utterance u_u.

The dialogue system is integrated in an event-driven, blackboard architecture (Buckley and Benzmüller, 2007) revolving around the shared belief state $\mathbf{b}$ represented as a Bayesian Network. This belief state is read and written by all the dialogue models. Once a change occurs on a state variable, the algorithm checks whether there are models triggered by this update. Then, for each triggered model, the rules are applied as follows:

1. For every rule r in the model, we create a rule node ϕ_r and include the conditional dependencies with its input variables. If the rule is a probability rule, the rule node will be a chance node describing the distribution of effects given the input assignment. If the rule is a utility rule, the node will be a utility node describing the utility of action variables.

2. The nodes corresponding to the output variables are created (if they do not already exist). For probability rules, these nodes will be chance nodes with a conditional dependence on the rule node ϕ_r. For utility rules, they will be action nodes, with an outward dependence relation to the rule node ϕ_r.

Once no more models can be triggered, the Bayesian Network is modified to replace the updated variables. Finally, if the network contains action variables, the algorithm searches for their optimal action value and selects them. This selection might trigger other inference steps, and the process is repeated until stability is reached. The procedure is described in Algorithms 1 and 2. Figure 3 illustrates the application of four rules on a belief state.

Once the Bayesian network is updated with the new rules, queries can be evaluated using any standard algorithm for exact or approximate inference.

Algorithm 1 : BELIEFUPDATE $(\mathbf{b}, \mathcal{M})$

Require: $\mathbf{b}$: Current belief state
Require: $\mathcal{M}$: Set of rule-based models
loop
  repeat
    for all model $m \in \mathcal{M}$ **do**
      if m is triggered **then**
        for all rule $r \in m$ **do**
          $\mathbf{b} \leftarrow \textsc{AddRule}(\mathbf{b}, r)$
        end for
      end if
    end for
  until no model is triggered
  for all node $x' \in \mathbf{b}$ **do**
    Prune x from $\mathbf{b}$
    Relabel x' into x
  end for
  for all node a : action variable **do**
    Find $a^* = \arg\max_{v \in Val(a)} Q(a{=}v|\mathbf{b})$
    Set $a \leftarrow a^*$
  end for
end loop

Algorithm 2 : ADDRULE $(\mathbf{b}, r)$

Require: $\mathbf{b}$: Current belief state
Require: r: Rule to add
$\mathcal{I}_r \leftarrow \textsc{InputVariables}(r)$
Create node $\phi_r \leftarrow \textsc{RuleNode}(r)$
Add ϕ_r and dependencies $\mathcal{I}_r \rightarrow \phi_r$ to $\mathbf{b}$
if r is a probability rule **then**
  $\mathcal{O}_r \leftarrow \textsc{OutputVariables}(r)$
  for all output variable $o \in \mathcal{O}_r$ **do**
    Create node o' if not already in $\mathbf{b}$
    Add o' and dependency $\phi_r \rightarrow o'$ to $\mathbf{b}$
  end for
else if r is a utility rule **then**
  $\mathcal{A}_r \leftarrow \textsc{ActionVariables}(r)$
  for all action variable $a \in \mathcal{A}_r$ **do**
    Create node a' if not already in $\mathbf{b}$
    Add a' and dependency $a' \rightarrow \phi_r$ to $\mathbf{b}$
  end for
end if
return $\mathbf{b}$

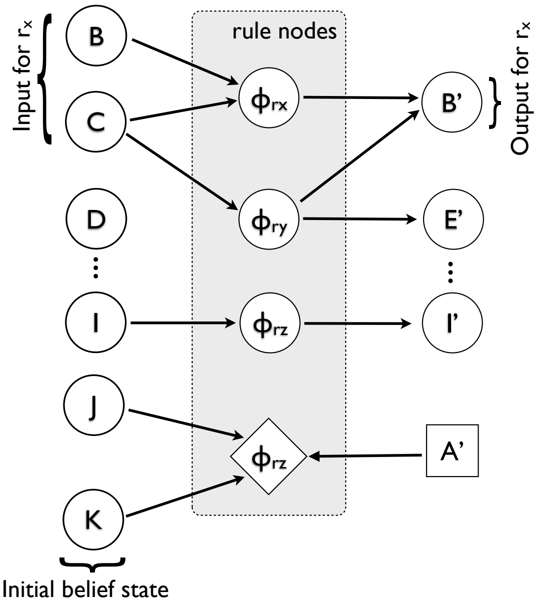


Figure 3: Bayesian Network expanded with the rules r_w, r_x, r_y, r_z on a set of state variables. Diamond nodes represent value nodes, and square action nodes.

It is worth nothing that the outlined procedure is an instance of *ground inference* (Getoor and Taskar, 2007), since the rule structure is grounded in a standard Bayesian Network.

4 Implementation

The framework outlined in the previous section has been implemented in a system architecture, and applied to a human-robot interaction scenario. The scenario involved a human user and a Nao robot[1] (see Figure 4). The user was instructed to teach the robot a sequence of basic movements (lift the left arm, step forward, kneel down, etc.) using spoken commands. The interaction included various dialogue acts such as clarification requests, feedbacks, acknowledgements, corrections, etc.

The models for speech understanding, action selection and generation were all encoded with probabilistic rules, for a total of 68 rules. The expressivity of the formalism allows us to capture complex probability or utility models in just a handful of rules.

[1] A programmable humanoid robot developed by Aldebaran Robotics, `http://www.aldebaran-robotics.com`.

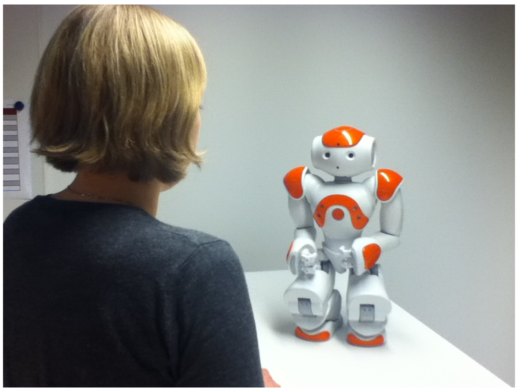


Figure 4: Human user interacting with the Nao robot to teach a sequence of movements via verbal instructions.

The structure of these rules is designed by hand, but their parameters can be learned from data. In our experiments, the utility parameters of the action selection module were estimated from limited amounts of Wizard-of-Oz training data, while the understanding and generation models were hand-crafted – these rules encoding simple, deterministic pattern matching techniques.

In addition, the dialogue also included a speech recognizer (Vocon 3200 from Nuance) connected to the robot microphones, a text-to-speech module, as well as components for planning the robot movements and controlling its motors in real-time. In our experiments, the posteriors for the updated state variables were calculated via importance sampling (Koller and Friedman, 2009), but other inference algorithms such as variable elimination (Zhang and Poole, 1996) are also available.

Both the rule-based models and the external modules are connected to the shared belief state, and read/write to it as they process their data flow. The models and external modules listen for changes occurring in the belief state and become activated when one of these changes relates to their triggering variable. It is worth noting that external modules such as the ASR engine exhibit the same processing behaviour as probabilistic rules, i.e. they extend the belief state with additional rule nodes that themselves lead to updated variables.

4.1 Example of rules

We describe below five examples of rules: two used for shallow dialogue act recognition, two used for action selection, and one for generation.

1. Rule r_1 below lists three regular expression patterns for a particular case of dialogue act classification. If the value for the user utterance variable u_u matches at least one of the patterns, the dialogue act a_u is classified as LeftArmDown, else a_u is left unchanged:

$$\begin{aligned} r_1 : \ & \textbf{if } (u_u \text{ matches "left arm down"}) \\ & \vee (u_u \text{ matches "lower * left arm"}) \\ & \vee (u_u \text{ matches "down * left arm"}) \ \textbf{then} \\ & \quad \{P(a'_u = \text{LeftArmDown}) = 1.0\} \end{aligned}$$

2. Due to high levels of noise in speech recognition, it is often useful to dynamically "prime" the results given expectations from the context. Rule r_2 is another rule for dialogue act recognition triggered if the last system act is AskRepeat. In this case, the rule r_2 will prime the probability that the new user act (a'_u) is identical to the previous one (a_u):

$$\begin{aligned} r_2 : \ & \textbf{if } (a_m = \text{AskRepeat}) \ \textbf{then} \\ & \quad \{P(a'_u = a_u) = 0.9\} \end{aligned}$$

Rule r_2 specifies that, if requested to repeat his last dialogue act, the user will do so with probability 0.9. The rule provides a *prediction* of the next user act given the context, before the observation of the user utterance. In combination with classification rules such as r_1, the rule determines the posterior distribution over the most likely dialogue acts uttered by the user.

3. Rule r_3 is an action selection rule which specifies the utility of performing the action $a_m = \text{DoMovement(X)}$ if the user intention i_u is equal to $\text{RequestMovement(X)}$, where X is an argument corresponding to the actual movement (lifting the arm up or down, etc.):

$$\begin{aligned} r_3 : \ & \textbf{if } (i_u = \text{RequestMovement(X)}) \ \textbf{then} \\ & \quad \{Q(a'_m = \text{DoMovement(X)}) = 3.0\} \end{aligned}$$

Note the use of the unbounded variable X, which is unified at runtime with the actual argument value for i_u.

4. Rule r_4 specifies the utility of the clarification request $a_m = \text{AskRepeat}$. The rule r_4 has no condition, which means that the utility of the clarification request will be conditionally independent of the value of i_u:

$$r_4: \textbf{if } (true) \textbf{ then} \\ \{Q(a'_m = \text{AskRepeat}) = 1.2\}$$

   Put together, rules r_3 and r_4 determine the relative utility of requesting a clarification on the user intention vs. performing the action. In this particular case, the system will select DoMovement(X) if the user intention $\text{RequestMovement(X)}$ has a probability $>$ 0.4, and will ask for a clarification otherwise.

5. Finally, rule r_5 determines the system utterance to synthesise u_m given the system act $a_m = \text{Ack}$ (for "acknowledgement"). In this case, the system is free to select one of the three alternatives, with equal utility:

$$r_5: \textbf{if } (a_m = \text{Ack}) \textbf{ then} \\ \{Q(u'_m = \text{"ok"}) = 1.0 \ \wedge \\ Q(u'_m = \text{"great"}) = 1.0 \ \wedge \\ Q(u'_m = \text{"thanks"}) = 1.0\}$$

For the sake of simplicity, the probability and utility values shown above were hand-coded. Of course, dialogue systems deployed in real domains need to estimate these parameters from interaction data (coming from e.g. Wizard-of-Oz experiments). Previous work has demonstrated how to perform such parameter estimation using a Bayesian learning approach (Lison, 2012). One major benefit is that the rule structure is described in exponentially fewer parameters than its plain counterpart, and is thus much easier to learn and to generalise to unseen data.

It should be theoretically possible to also learn the rule *structure* from data, as evidenced by work done in Statistical Relational Learning (Pasula et al., 2007). Such endeavour would however require significantly larger amounts of training data, and remains therefore impractical for most dialogue domains. Furthermore, the rule structure can be seen as a way for the system designer to enforce *design constraints* or business rules into the system (Williams, 2008), and such ability would be lost if the rule structure was to be learned from scratch.

5 Discussion and related work

The development of generic, domain-independent dialogue systems has a long history (Allen et al., 2000; Bohus et al., 2007), and there is a clear trend towards creating platforms with generic or reusable components. There is however no agreement on common modelling formats or processing techniques. In this respect, it is interesting to draw a parallel between dialogue systems and other fields of NLP such as syntactic parsing. Before the 60's, most parsers relied on procedural routines buried in the code. One of the major advances has come from the decision to separate the domain knowledge (in this case, the lexicon and grammar) on one hand, and the parsing algorithms on the other hand. We believe that dialogue systems would also benefit from a cleaner distinction between declarative knowledge (i.e. task- and domain-specific models) and generic processing functionalities (i.e. algorithms for reasoning, learning and planning under uncertainty.)

Most current dialogue systems are however relying on numerous blackbox components in their pipeline. An unfortunate consequence of this heterogeneity is that, while speech recognition results often include explicit measures of uncertainty (in the form of e.g. confidence scores), this uncertainty is often lost at higher stages of reasoning, such as semantic interpretation and dialogue management. Recent papers have shown that confidence scores can be exploited in dialogue management (Williams et al., 2008), but their approach has not yet been widely adopted. Thanks to the unified description framework provided by probabilistic rules, our approach is able to provide a principled account for this uncertainty at all processing stages.

Information-state approaches to dialogue management (Larsson and Traum, 2000; Bos et al., 2003) are closely related to this work, since they also rely on a shared state updated according to a rich repository of rules, but contrary to the approach presented here, these rules are generally deterministic and do not include learnable parameters. The idea of state space partitioning, implemented here via rule conditions, has also been explored in recent papers (Williams, 2010; Crook and Lemon, 2010).

The work presented in this paper can be seen as an attempt to bridge the gap between "symbolic" approaches to dialogue, which usually concentrate on capturing rich interaction patterns, and "probabilistic" approaches, more focused on aspects related to noise, uncertainty, and learnability. There has been some initial work on hybrid approaches to dialogue processing and management where both statistically learned and designed policies are combined (Williams, 2008; Lee et al., 2010), but they generally use the designed policies as a mere filtering mechanism for the stochastic policy. Our approach however directly incorporates the prior domain knowledge into the statistical model.

Structural knowledge in probabilistic models has been explored in many directions, both in decision-theoretic planning and reinforcement learning (Hauskrecht et al., 1998; Pineau, 2004; Lang and Toussaint, 2010; Otterlo, 2012) and in statistical relational learning (Jaeger, 2001; Richardson and Domingos, 2006; Getoor and Taskar, 2007). The introduced structure may be hierarchical, relational, or both. As in our approach, most of these frameworks rely on expressive representations as *templates* for grounded probabilistic models.

An important side benefit of structured representations in probabilistic models is their improved readability for the human designers, which are able to use these powerful abstractions to encode their prior knowledge of the dialogue domain in the form of pragmatic rules, generic background knowledge, or task-specific constraints. Moreover, the grouping of related rules into models allows the system developer to specify dialogue domains in a modular fashion, by clustering rules into various sets of models. Some models might be highly domain-specific while others encode generic interaction behaviours that can be easily ported to other applications.

6 Conclusion

We have described in this paper a new approach to the development of dialogue systems, based on the declarative specification of *probabilistic rules*. These rules are defined as structured mappings over variables of the dialogue state, specified using high-level conditions and effects. The rules are parametrised with effect probabilities or action utilities. Probabilistic rules allow the system designer to exploit powerful generalisations in the dialogue domain specification without sacrificing the probabilistic nature of the model.

The architecture revolves around a shared belief state expressed as a Bayesian Network. This belief state is continuously updated and extended based on a set of probabilistic rules for speech understanding, management and generation. This architecture has been implemented and integrated in a spoken dialogue system for human-robot interaction. We are currently in the process of refactoring our implementation to make it available as a generic, open source dialogue toolkit called openDial.

We are currently working on extending this architecture in several directions. Our first line of work is to extend the parameter estimation outlined in (Lison, 2012) to Bayesian model-based reinforcement learning. The parameter estimation currently operates in a supervised learning mode, which requires expert data. Alternatively, one could estimate the model parameters in a fully online fashion, without any supervisory input, by incorporating model uncertainty into the inference and continuously adapting the parameter distribution from real or simulated interaction experience (Ross et al., 2011).

Another research direction relates to the extension of the belief update algorithms towards incrementality (Schlangen et al., 2010). We believe that the framework presented in this paper is particularly well suited to perform incremental processing, since the chain of related hypotheses is explicitly captured in the conditional dependencies of the Bayesian Network. A probability change in one initial hypothesis (e.g. the user utterance) will therefore be directly reflected in all hypotheses depending on it (e.g. the corresponding user intention). Extending the belief update algorithm to run incrementally while remaining tractable is however a non-trivial task.

References

J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering*, 6:213–228.

D. Bohus, A. Raux, T. K. Harris, M. Eskenazi, and A. I.

Rudnicky. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, NAACL-HLT-Dialog '07, pages 32–39, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Bos, E. Klein, O. Lemon, and T. Oka. 2003. DIPPER: Description and formalisation of an information-state update dialogue system architecture. In *4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124.

M. Buckley and C. Benzmüller. 2007. An agent-based architecture for dialogue systems. In *Proceedings of the 6th international Andrei Ershov memorial conference on Perspectives of systems informatics*, PSI'06, pages 135–147, Berlin, Heidelberg. Springer-Verlag.

P. A. Crook and O. Lemon. 2010. Representing uncertainty about complex user goals in statistical dialogue systems. In *Proceedings of the 11th SIGDIAL meeting on Discourse and Dialogue*, pages 209–212.

L. Getoor and B. Taskar. 2007. *Introduction to Statistical Relational Learning*. The MIT Press.

M. Hauskrecht, N. Meuleau, L. P. Kaelbling, T. Dean, and C. Boutilier. 1998. Hierarchical solution of markov decision processes using macro-actions. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 220–229.

M. Jaeger. 2001. Complex probabilistic modeling with recursive relational bayesian networks. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):179–220.

D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

G.-J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings from the Symposium*, Aveiro, Portugal, 12.

T. Lang and M. Toussaint. 2010. Planning with noisy probabilistic relational rules. *Journal of Artificial Intelligence Research*, 39:1–49.

S. Larsson and D. R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natuarl Language Engineering*, 6(3-4):323–340, September.

C. Lee, S. Jung, K. Kim, and G. Geunbae Lee. 2010. Hybrid approach to robust dialog management using agenda and dialog examples. *Computer Speech & Language*, 24(4):609 – 631.

P. Lison. 2012. Probabilistic dialogue models with prior domain knowledge. In *Proceedings of the SIGDIAL 2012 Conference*, pages 179–188, Seoul, South Korea, July.

M. Otterlo. 2012. Solving relational and first-order logical markov decision processes: A survey. In *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pages 253–292. Springer Berlin Heidelberg.

H. M. Pasula, L. S. Zettlemoyer, and L.P. Kaelbling. 2007. Learning symbolic models of stochastic domains. *Journal of Artificial Intelligence Research (JAIR)*, 29(1):309–352, July.

J. Pineau. 2004. *Tractable Planning Under Uncertainty: Exploiting Structure*. Ph.D. thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, USA.

M. Richardson and P. Domingos. 2006. Markov logic networks. *Machine Learning*, 62:107–136.

S. Ross, J. Pineau, B. Chaib-draa, and P. Kreitmann. 2011. A Bayesian Approach for Learning and Planning in Partially Observable Markov Decision Processes. *Journal of Machine Learning Research*, 12:1729–1770.

D. Schlangen, T. Baumann, H. Buschmeier, O. Buß, S. Kopp, G. Skantze, and R. Yaghoubzadeh. 2010. Middleware for Incremental Processing in Conversational Agents. In *Proceedings of the 11th SIGDIAL meeting on Discourse and Dialogue*.

V. Thomson and S. Young. 2010. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Computer Speech & Language*, 24:562–588, October.

J. D. Williams and S. Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21:393–422.

J. D. Williams, P. Poupart, and S. Young. 2008. Partially observable markov decision processes with continuous observations for dialogue management. In Laila Dybkjr and Wolfgang Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*, pages 191–217. Springer Netherlands.

J. D. Williams. 2008. The best of both worlds: Unifying conventional dialog systems and POMDPs. In *International Conference on Speech and Language Processing (ICSLP 2008)*, Brisbane, Australia.

J. D. Williams. 2010. Incremental partition recombination for efficient tracking of multiple dialog states. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5382–5385.

N. Lianwen Zhang and D. Poole. 1996. Exploiting causal independence in bayesian network inference. *Journal of Artificial Intelligence Research (JAIR)*, 5:301–328.

Communicating with Cost-based Implicature: a Game-Theoretic Approach to Ambiguity

Hannah Rohde
University of Edinburgh
3 Charles Street
Edinburgh, UK
hannah.rohde@ed.ac.uk

Scott Seyfarth
UC San Diego
9500 Gilman Drive
La Jolla, CA, USA
sseyfarth@ucsd.edu

Brady Clark
Northwestern University
2016 Sheridan Road
Evanston, IL, USA
bzack@northwestern.edu

Gerhard Jaeger
University of Tübingen
Wilhelmstraße 19
Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Stefan Kaufmann
Northwestern University
2016 Sheridan Road
Evanston, IL, USA
kaufmann@northwestern.edu

Abstract

A game-theoretic approach to linguistic communication predicts that speakers can meaningfully use ambiguous forms in a discourse context in which only one of several available referents has a costly unambiguous form and in which rational interlocutors share knowledge of production costs. If a speaker produces a low-cost ambiguous form to avoid using the high-cost unambiguous form, a rational listener will infer that the high-cost entity was the intended entity, or else the speaker would not have risked ambiguity. We report data from two studies in which pairs of speakers show alignment of their use of ambiguous forms based on this kind of shared knowledge. These results extend the analysis of cost-based pragmatic inferencing beyond that previously associated only with fixed lexical hosts.

1 Introduction

A growing body of work demonstrates that joint communication tasks yield alignment of referring expressions, highlighting the role of interlocutors' experience of shared common ground in establishing conventions (Brennan & Clark, 1996; Garrod & Pickering, 2004). Less well-established, however, are predictions regarding which form∼meaning mappings interlocutors will converge on. To address this, we evaluate alignment in contexts where interlocutors' common ground includes the costs of producing particular forms.

Consider the shapes in Figure 1. In a context that contains only the first item, a speaker can efficiently draw attention to it by saying "Look at the circle." In a context with all three shapes, a more specific referring expression—such as "blue circle"—is required to unambiguously indicate that same item. However, if it is necessary to draw attention to the third item, the speaker may need to accept either inefficiency or ambiguity. Since there is no efficient label (e.g., "circle") for the third item's unique shape, it is costly to unambiguously refer to it in the context of Figure 1—a longer or more obscure expression is necessary (e.g., "the triangle-and-square thing" or "the blue shape that's not a circle"). On the other hand, the speaker can avoid producing a costly expression by instead using an ambiguous expression such as "the blue thing" or "the blue shape." The question is whether a listener can be expected to infer that the intended referent is the difficult-to-describe shape, even though "the blue shape" could in principle also refer to the blue circle.

An accurate inference about the intended referent of "the blue shape" requires the following chain of reasoning: The listener would have to realize that had the speaker intended to refer to the blue circle, a relatively short unambiguous expression would have sufficed ("the blue circle"); since the speaker used a low-cost ambiguous expression "the blue shape" instead of the available low-cost unambiguous name,

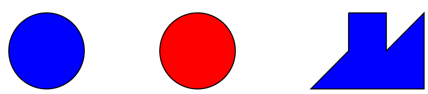


Figure 1: Variation in cost of referring expressions

the implicature is that the circle is not the intended referent, leaving the difficult-to-describe item as the preferred target.

Using ambiguous forms to convey meaning depends in part on the listener's ability to diagnose the source of the ambiguity: Does the ambiguity likely arise from the speaker's own production decisions or from other factors that make the expression noisy or unclear (see Schlangen & Fernandez, 2007)? We focus here on productions whose ambiguity listeners can be confident originates with the speaker herself.

This paper presents two studies testing speaker alignment in dialog games with superimposed costs and rewards that are predicted to guide production and comprehension of ambiguous forms. Our results show that speakers' use of ambiguous expressions reflects the relative costs of available forms. Rather than avoiding ambiguity, speakers show behavior that is in keeping with theories of communicative efficiency that posit that speakers make rational decisions about redundancy and reduction.

2 Game Theory and Implicature

Game theory is an area of applied mathematics, which aims to provide a framework for analyzing the behavior of individuals (*players*) in strategic situations (*games*) (von Neumann and Morgenstern, 1944; see Clark, 2011, Chapter 4 for an introduction). It is used to describe games in which players have choices regarding their behavior and preferences over possible outcomes. The outcomes depend on both players' choices: While some games are zero-sum, meaning that success requires that only one player can win, in other types of games, both players can succeed if they coordinate their actions. Linguistic communication is typically argued to be an example of this second type (Lewis, 1969).

Games are characterized by shared knowledge, meaning that players know the overall structure of the game. They know what moves are possible in which situation by each player, what consequences are associated with each move, and what preferences each player has. Crucially, all players know that the other players know these facts. A game-theoretic approach makes an assumption of player rationality.

In a recent computational model that treats language use as a cooperative game, Golland, Liang, and Klein (2010) show that a rational speaker's decisions about whether or not to use ambiguous referring expressions can be significantly improved by embedding a model of the listener that represents the shared game knowledge described above. In their communication game with a human listener, unambiguous expressions were preferred over ambiguous-yet-accurate expressions when the speaker selected an expression that optimized utility for the listener.

However, in that type of listener-oriented model, a speaker's choice of expression is based solely on maximizing the probability that the listener understands the speaker's intent. In this paper, we consider how a game-theoretic approach offers further predictions about the players' behavior when the various (ambiguous and unambiguous) referring expressions also have different costs, when the players share knowledge of costs, and when the players know that they share knowledge. In particular, this approach suggests cases in which literal ambiguity might actually be preferred in pursuit of efficient communication.

The prediction from such an approach is that an ambiguous form can be used to refer unambiguously if an unambiguous form is costly and other meanings can be conveyed at low cost (Clark, 2011; Jäger, 2008). In other words, a listener who knows the relative costs of unambiguously referring to X (high-cost) or Y (low-cost) may reason that a speaker using an ambiguous word "X-or-Y" (low-cost) intends to convey X, or else she would have said "Y".

This type of reasoning has been used to explain the conventional use of "some" (Jäger, 2008). Having heard a speaker use the word "some", the listener is faced with two possible interpretations: AT-LEAST-ONE-AND-POSSIBLY-ALL or else AT-LEAST-ONE-AND-NOT-ALL. A rational listener is said to reason that, had the speaker intended to convey the meaning ALL, she would have used the low-cost (short and easy to produce) form "all". The speaker, having instead used a low-cost but ambiguous form "some", can be taken to implicate that the intended interpretation is not ALL, but is instead a meaning that would have been costly to produce unambiguously: AT-LEAST-ONE-BUT-NOT-ALL.

This account of "some" formulates in game-theoretic terms the well-known Gricean account, which focuses on the amount of information conveyed. In the Gricean version, the literal meaning of "some", AT-LEAST-ONE-AND-POSSIBLY-ALL, conveys less information than "all". Its meaning is strengthened to a more informative meaning of AT-

LEAST-ONE-BUT-NOT-ALL via implicature: A cooperative speaker who obeys the maxim of Quantity and intends to convey the more informative meaning ALL would have said "all", but since she did not, the meaning AT-LEAST-ONE-BUT-NOT-ALL is favored.[1] Grice's recognition of the importance of speaker intention echoes a game-theoretic approach to signaling and the calculation of what must be true in order for a rational speaker to have produced a particular utterance (Stalnaker, 2005).

The AT-LEAST-ONE-BUT-NOT-ALL implicature associated with the use of "some" is what Grice called a generalized conversational implicature: the implicature AT-LEAST-ONE-BUT-NOT-ALL is typically associated with the proposition expressed. What remains an open question is whether this type of cost-based inferencing applies beyond a fixed lexical host like "some". The next sections describe two studies aimed at measuring alignment in a communication game with explicit superimposed costs and rewards for production and comprehension.

3 Study 1: Communicating about Objects with Divergent Costs

We created a networked interactive two-player chat environment (see Figure 2) in which pairs of players could communicate about a set of objects. Costs and rewards were made explicit via points, and players shared both knowledge of the cost/reward structure as well as a shared goal of working together to communicate successfully. In contrast to Study 2 in the next section, the costs imposed in Study 1 served to highlight a single high-cost entity in each category, creating a bias to conventionalize the meaning of an ambiguous form to refer to that entity. In production, the prediction is that players will use a low-cost *ambiguous* word to refer to an object with a high-cost *unambiguous* name, as long as other objects can be unambiguously referred to with relatively low-cost names. In comprehension, the prediction is that players will more often interpret ambiguous words to refer to objects with a costly unambiguous name than to objects whose unambiguous name is associated with a mid or low cost (henceforth 'high-cost objects', 'mid-cost objects', and 'low-cost objects').

3.1 Participants

10 pairs of English speakers from Northwestern University received $10 to participate in the study.

3.2 Methods

Materials The game involved a set of objects in two categories—three flowers and three trees. Players could communicate using a set of eight referring expressions: six unambiguous names and two ambiguous words. The costs varied among the unambiguous names, but the ambiguous words were both low-cost. Table 1 shows the point costs associated with the eight forms. The point values themselves are less important than the relative values: In this study, the cost of the most expensive name in each category ("Tulip"/"Pine Tree") was more than four times the cost of the least expensive name and more than twice the cost of the mid-cost name.[2]

Name	*Cost*	*Name*	*Cost*
"Rose"	-60	"Apple Tree"	-60
"Daisy"	-120	"Palm Tree"	-120
"Tulip"	-280	"Pine Tree"	-250
"Flower	-80	"Tree"	-80

Table 1: Referring expressions and their costs (Study 1)

Task Players were seated in separate rooms in front of computers showing the interactive game interface depicted in Figure 2. They were told that they would take turns as Sender and Receiver in a game that involved communicating about a set of objects. As the Sender, a player would see a gnome highlight an object with a spotlight, and the Sender's task was to send a message to the other player so that the other player (the Receiver) could guess what the highlighted object must have been. Sending a word consisted of pressing a button on the screen which

[1] This logic is spelled out in the Stanford Encyclopedia of Philosophy (Davis, 2010) in terms of the interaction of cost (maxim of Manner) and information (maxim of Quantity):

> Assuming that the accepted purpose of the conversation requires the speaker to say whether or not all athletes smoke, a speaker who said "Some athletes smoke" would be infringing the Quantity maxim if she meant only what she said. So she must have meant more. If she believed that all athletes smoke, she would have said so. Since she did not, she must have meant that some but not all athletes smoke. As a bonus, she achieved brevity, in conformity to the maxim of Manner.

[2] Alternate materials were constructed with abstract shapes and nonsense names, but a pilot study found that participants had difficulty learning the names. Variants of Study 1 were conducted with first names (e.g., "Ann") for unambiguous names for objects in plant and vessel categories and family names or nonsense names (e.g., "Puliniki") for ambiguous words; the results matched those presented here with flower/tree materials.

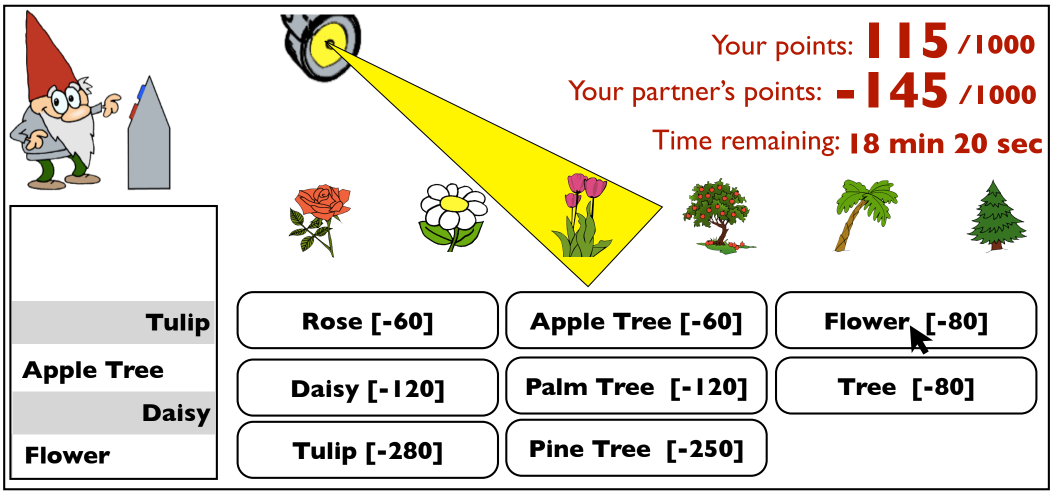


Figure 2: Player's view of the interactive chat environment for both Study 1 and 2 (point values from Study 1)

displayed the name and cost; pressing the button resulted in an immediate point decrement for the Sender. If the Receiver successfully identified the intended object, then both players got an immediate reward (Sender and Receiver scores increased). Otherwise, there was no reward and the Sender had to retry until communication succeeded—the penalty in that case being the continued point decrements for sending multiple words. The reward for successful communication was +85 points for each player. If either player reached the target score of 1000 points, the game ended and both players were free to leave. Otherwise, games continued for 20 minutes or until the gnome had highlighted a total of 60 objects. Scores could become negative if the point decrements of production outstripped the point increments for successful communication.

For each pair, we calculated a unique sequence of objects to be highlighted so that it would be impossible to reach 1000 points in less than 60 turns without successfully coordinating the use of ambiguous words. The first 10 turns were intended as practice turns, involving only low-cost and mid-cost objects.

Shared Knowledge Both players were able to see the available referring expressions and their costs and were told that the other player could as well. Both players were informed of the reward structure (+85 points each for successful communication, 1000 points to end early) and told that the other player had likewise been informed. The game interface showed the current scores of the two players, the time remaining, and a scrolling chat window listing previous words sent and received.

3.3 Results and Discussion

Of the 10 pairs, 5 consistently coordinated their referring expressions, allowing an early exit from the game. Two pairs played for the full 20 minutes, struggling to coordinate their efforts, and their limited attempts at coordination failed to yield an early exit to the game. Of the remaining 3 pairs, 1 pair did not make a serious attempt at using the ambiguous words, and the other 2 pairs used them repeatedly but had difficulty finding a strategy while doing so, although all 3 did eventually manage to exit the game before reaching the time or turn limit.

Table 2 shows a transcript from one pair of players, listing the first 26 moves of their game. The transcript demonstrates how the players developed a coordinated strategy for using the ambiguous words: the use of an ambiguous word by Player 1 when the intended referent was not a high-cost object (which led to Player 2's initial guess that the high-cost object was the target), the use of an ambiguous word by Player 2 (which Player 1 failed to interpret as a reference to the high-cost object), and eventually their convergence. After the success shown in the last line of the table, the players continued to reliably use "Flower" and "Tree" to refer to the tulip and pine tree, and the game ended after 44 moves when Player 2 reached 1000 points.

For the analysis, we measured the effect of one within-players factor (the target object's unambiguous cost) on two binary outcomes: whether the Sender sent an ambiguous word (production outcome) and whether an ambiguous word resulted in successful communication (comprehension out-

Sender	*Highlighted Target*	*Form Used*	*Receiver's Guess*
1	daisy (mid)	'Flower'	tulip (high)
1	daisy (mid)	'Daisy'	daisy (mid)
2	palm (mid)	'Palm'	palm (mid)
1	palm (mid)	'Tree'	pine (high)
1	palm (mid)	'Palm'	palm (mid)
2	apple (low)	'Apple'	apple (low)
1	palm (mid)	'Tree'	pine (high)
1	palm (mid)	'Palm'	palm (mid)
2	daisy (mid)	'Daisy'	daisy (mid)
1	tulip (high)	'Tulip'	tulip (high)
2	apple (low)	'Apple'	apple (low)
1	pine (high)	'Pine'	pine (high)
2	tulip (high)	'Flower'	daisy (mid)
2	tulip (high)	'Flower'	daisy (mid)
2	tulip (high)	'Flower'	tulip (high)
1	pine (high)	'Pine'	pine (high)
2	pine (high)	'Tree'	pine (high)
1	rose (low)	'Rose'	rose (low)
2	rose (low)	'Rose'	rose (low)
1	palm (mid)	'Palm'	palm (mid)
2	tulip (high)	'Flower'	rose (low)
2	tulip (high)	'Flower'	daisy (mid)
2	tulip (high)	'Flower'	tulip (high)
1	pine (high)	'Tree'	apple (low)
1	pine (high)	'Tree'	pine (high)
2	pine (high)	'Tree'	pine (high)
⋮	⋮	⋮	⋮

Table 2: Excerpt of a game transcript from two successful players in Study 1 ('Apple Tree', 'Palm Tree', and 'Pine Tree' are abbreviated without the word 'Tree'). Consecutive rows with the same Sender show retries.

come). We also measured the effect of trial number on the interpretation of ambiguous words in order to evaluate the time course of the Receivers' cost-based inferencing. For that, the three-way outcome of Receiver guess was treated as three binary variables (high-cost-or-not, mid-cost-or-not, low-cost-or-not). For all analyses, means, and figures, we consider only non-retry moves. We report the logistic-regression coefficient estimate and p-value (based on the Wald Z statistic; Agresti, 2002) for the factors *cost* and *trial number* (both treated as numeric factors) with random participant-specific intercepts and slopes.

Additionally, Figure 3 shows the overall rates of

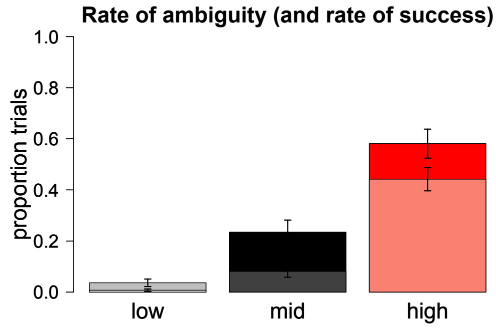


Figure 3: The percentage of trials with a low-, mid-, or high-cost target in which a Sender produced an ambiguous word. The lower part of each bar represents the percentage of trials in which the use of an ambiguous word was successful.

use and success for the ambiguous words, measured over proportions of trials. The height of each bar in the graph shows the proportion of trials for each cost—low, mid, high—where an ambiguous word was used. The shaded (lower) portion of the bar shows the proportion of those ambiguous words that resulted in successful communication.

As predicted, Senders produced an ambiguous word more often if the gnome had highlighted a target object whose unambiguous name was high cost (β_{cost}=1.94, p<0.001): 58.9% of high-cost-target trials yielded an ambiguous word, whereas only 24.8% of mid-cost-target trials and 4.6% of low-cost-target trials yielded ambiguity.

Receivers likewise paid attention to cost, correctly guessing the target more often when an ambiguous word was used for a high-cost target than for a mid- or low-cost target (β_{cost}=1.68, p<0.005): Trials with an ambiguous word yielded successful communication 74.1% of the time if the intended target was high cost, compared to 37.8% and 37.5% success when an ambiguous word was used for mid-cost and low-cost targets, respectively. In other words, of the 58.9% of trials in which an ambiguous word was produced for a high-cost target, 74.1% of those trials yielded successful communication (compared to less than half the time for trials in which an ambiguous word was used for a low- or mid-cost target), as depicted in the 'high' bar of Figure 3.

We also ask whether the interpretation of an ambiguous word changes over successive trials. Restricting the analysis to trials in which an ambiguous word was used, we find that the interpreta-

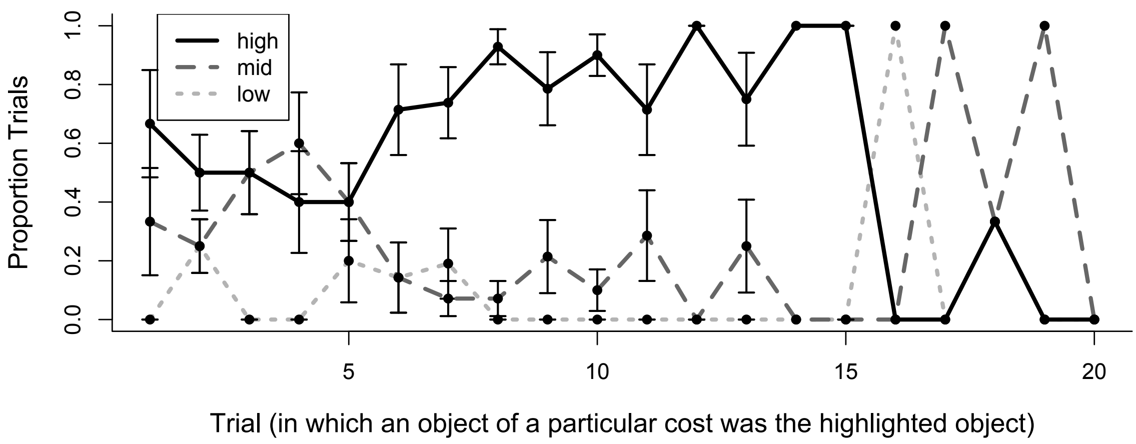


Figure 4: Rate at which Receivers in Study 1 interpreted ambiguous words as referring to objects with high-, mid-, and low-cost unambiguous names. Errors bars show standard error of the mean.

tion of ambiguous words favors high-cost objects over time (β_{trial}=0.29, p<0.01) and disfavors, albeit not significantly, mid-cost and low-cost objects (mid: β_{trial}=-0.22, p=0.09; low: β_{trial}=-0.23, p=0.11). Figure 4 shows the probability of a Receiver interpreting an ambiguous word as referring to a high, mid, or low-cost object. Trial number in Figure 4 (and in the regression) represents the number of trials for which an object of that cost has been highlighted—e.g., the datapoints at Trial=10 are the 10th trials, within each game, in which the gnome highlighted a high-cost object (either a tulip or a pine tree) and the Sender sent an ambiguous word (‘Flower’ or ‘Tree’) and the Receiver guessed a high-cost object (solid line), a mid-cost object (dashed line), or a low-cost object (dotted line).

What is apparent in Figure 4 is that the data after Trial=15 becomes noisier (more fluctuation) and more sparse (limited/no error bars). This drop-off corresponds to the point in the game when most successful players had reached 1000 points and left, so the data for the later trials represents the behavior of pairs of players who had failed to coordinate their referring expressions. These players overall used fewer ambiguous words, and because of this, many were watching their scores become more and more negative. Data from all players up through Trial=15 was analyzed in the time course logistic regression.

These results show that players can and do coordinate their use of referring expressions, conventionalizing the meaning of an ambiguous form to associate it with the object whose unambiguous name is the most costly to produce.

4 Study 2: Communicating about Objects with Similar Costs

As a further test of the predictions of a game-theoretic model, a second study was conducted that shifted the cost structure within each category. Compared with the costs in Study 1, the high-cost names in Study 2 were less costly than before, and the point difference between the low-cost, mid-cost, and high-cost names was smaller. The revised costs were predicted to reduce the likelihood that players would coordinate their use of referring expressions. The target score and the reward for successful communication stayed the same, but the stakes were lower (i.e., the cost structure imposed lower costs overall) so that it was possible for players to reach the target score in less time without making recourse to the ambiguous words. Rational players could choose to waste less effort attempting to align their use of referring expressions because the benefit of alignment was potentially outweighed by the points lost in rounds in which successful communication required the Sender to send more than one word.

4.1 Participants

10 pairs of English speakers from Northwestern University received $10 to participate. None had participated in Study 1.

4.2 Methods

The game environment contained the same set of six objects. The game rules and shared knowledge of those rules matched Study 1. The only difference was the costs (Table 3): The most expensive name

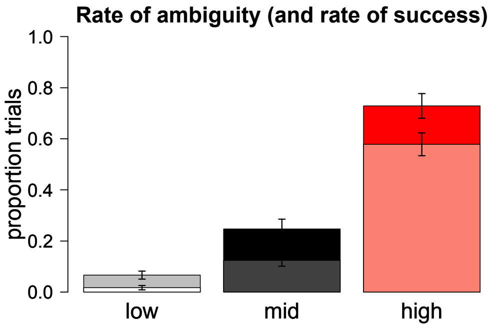


Figure 5: The percentage of trials with a low-, mid-, or high-cost target in which a Sender produced an ambiguous word. The lower part of each bar represents the percentage of trials in which the use of an ambiguous word was successful.

in each category cost slightly more than two times the least expensive name and not more than one and a half times the mid-cost name. Successful communication was still rewarded with +85 points to both players and the game ended when either player reached 1000 points. Since the absolute costs for the low- and mid-cost objects were similar to Study 1 while the absolute cost for the high-cost object was reduced, the average point cost in Study 2 was reduced and therefore it was possible to end the game after 48 turns rather than 60 without coordination.

Name	*Cost*	*Name*	*Cost*
"Rose"	-80	"Apple Tree"	-80
"Daisy"	-140	"Palm Tree"	-135
"Tulip"	-165	"Pine Tree"	-170
"Flower	-80	"Tree"	-80

Table 3: Referring expressions and their costs (Study 2)

4.3 Results and Discussion

Of the 10 pairs, 8 coordinated their referring expressions, allowing an early exit from the game. Contrary to prediction, the imposition of lower costs did not reduce players' motivation to conventionalize. This can be seen in reliable effects of cost on production and comprehension, as in Study 1. Figure 5 shows the overall rates of use and success for ambiguous words in Study 2.

Senders produced an ambiguous word most often if the highlighted object was high cost (β_{cost}=2.56, p<0.001): 72.6% of high-cost-target trials yielded an ambiguous word, whereas only 25.6% of mid-cost-target trials and 6.4% of low-cost-target trials yielded ambiguity. Receivers likewise paid attention to cost, correctly guessing the target more often when an ambiguous word was used for a high-cost target than for a mid- or low-cost target (β_{cost}=1.18, p<0.001): Trials with an ambiguous word yielded successful communication 82.5% of the time if the intended target was high cost, compared to 51.0% and 21.4% success when an ambiguous word was used for mid-cost and low-cost targets, respectively.

The time course differs slightly from Study 1, however. Receivers in Study 2 did not show a reliable rise in their preference to interpret ambiguous words as referring to high-cost targets. Again restricting the analysis to trials in which an ambiguous word was used (see Figure 6), the only reliable effect is that ambiguous words become less likely to be interpreted as referring to low-cost objects over time (β_{trial}=-0.65, p=0.05). The effect of trial number on the use of ambiguity for mid-cost targets is again not significant, though the coefficient is positive here (it was negative in Study 1), meaning that ambiguity tended to favor the mid-cost target slightly over time (β_{trial}=0.13, p=0.32). The slight increase here in the rate at which ambiguous words were interpreted to refer to high-cost objects is not significant (β_{trial}=0.09, p=0.22), unlike in Study 1. Given the different cost structure, the point in the game when most successful players had reached 1000 points and left comes at Trial=11. Figure 6 shows the subsequent sparseness after that point, and the time course regression includes data only up to Trial=11.

Across the two studies, Sender/Receiver pairs who coordinated their use of ambiguous forms were better able to communicate. Two pairs in Study 2 converged on a pattern of usage whereby an ambiguous word was used to refer to the object with the mid-cost unambiguous name. This could explain the time course result whereby the slope for mid-cost guesses for ambiguous words was positive (though not significantly so) in this study but not in Study 1. Another difference between the two studies is that in Study 2, convergence in the use of ambiguity in one category did not guarantee convergence in the other: two pairs used 'Tree' reliably but not 'Flower'.

In terms of our predictions, players did show sensitivity to the differences in the cost structure, but what characterized the behavior of players in Study 2 was a more immediate and sustained use of ambiguous words as referring to high-cost objects for most pairs and an openness to assign the ambigu-

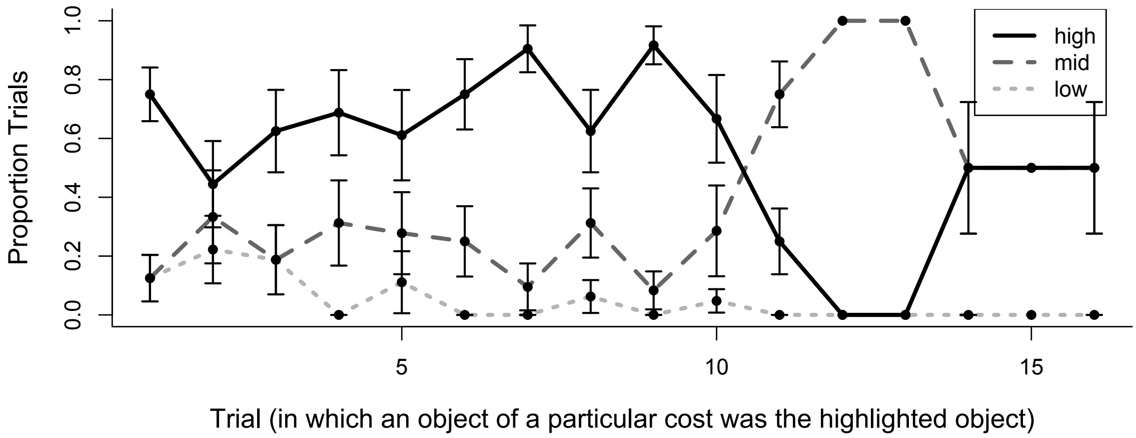


Figure 6: Rate at which Receivers in Study 2 interpreted ambiguous words as referring to objects with high-, mid-, and low-cost unambiguous names. Interpretation of axes and error bars as in Figure 3.

ous word to a mid-cost object for a small subset of pairs. The players' behavior seems to point to a greater willingness to experiment with the ambiguous words in a context like Study 2 where, despite their experimentation, they could see their scores increasing more quickly to the target value. Alternatively, rather than casting Study 2 as the context with increased experimentation, one can ask why players did not experiment more in Study 1. Perhaps the higher production costs in Study 1 made players avoid risking ambiguity and possibly having to retry.

Lastly, one can ask if players simply used a trial-and-error strategy for finding an efficient alignment without recourse to the pragmatic inference required for the cost-based implicature. To rule this out, we considered the trials in which a Sender *first* sent an ambiguous word. We pooled the data from the two studies since each participant could only contribute one datapoint. In keeping with the cost-based implicature account, Receivers inferred, more often than chance, that the high-cost object was the intended object ($\chi^2(2)$=11.54, p<0.005).

5 General Discussion

In keeping with the game-theoretic prediction, we saw that ambiguous words can be used meaningfully to refer to entities with costly unambiguous names, crucially if other referents can be identified with low-cost unambiguous names. This extends the claim that listeners draw cost-based implicature beyond the case of a fixed lexical host like "some".

We also saw sensitivity to relative costs: In comparing the two studies, the trajectory for how ambiguous words were interpreted over time in Study 1 (where the unambiguous names had more divergent costs) differs from the trajectory in Study 2 (where the unambiguous names had more similar costs). Only in Study 2 did players ever assign the ambiguous word to a mid-cost item, and only in Study 2 did a pair of successful players coordinate their use of one ambiguous word but not both. Contrary to predictions, however, the lower stakes in Study 2 did not yield a reduction in the players' overall ability to coordinate referring expressions.

This line of research raises an important question about how production cost should be measured. For the studies here, we imposed our costs arbitrarily—i.e., we could just as well have assigned the cost of 'Rose' to 'Tulip' or to 'Daisy'. One could imagine instead a cost metric that reflects properties such as length (as in Figure 1) or the presence of non-native phonemes or other articulation-based complexity. Alternatively, there is evidence that frequency contributes to production difficulty: Speakers are slower to name objects with low-frequency names than high-frequency names (Oldenfield & Wingfield, 1965). Speakers also experience difficulty, as measured by their disfluency, when describing objects which have not yet been mentioned (Arnold & Tanenhaus, 2007), are unfamiliar or lack a name (Arnold, Kam, & Tanenhaus, 2007). To avoid the inference step of assessing what phonological/lexical/pragmatic properties speakers perceive as costly, the proof-of-concept studies presented here used externally imposed costs to test the use of cost-based implicatures. The next stage of this research will extend the experimental framework described here to the kinds of costs that are imposed naturally in regular conversation.

If cost does influence choice of referring expression, one must still ask whether its role is automatic or strategic (Horton, 2008). By presenting this work from the standpoint of calculable implicatures, we have framed the questions in strategic terms. The factors that guide speakers' strategic selection of referring expressions may depend not only on the costs associated with production (as we have shown here) but also on speakers' estimates of the costs and benefits of successful communication and of the degree of coordination between speaker and hearer (van Deemter, 2009).

It is also possible that our participants had a more automatic reaction to the salience of high-cost objects — maybe they just associated an ambiguous form with the most salient object of that category, where cost indicated salience. This scenario is compatible with a game-theoretic account — the reasoning being that it would be unnecessarily costly to refer to a prominent entity with a full name when a reduced or ambiguous form could be used.[3]

Lastly, these results fit with existing work on the role of reduction in communication — namely, work showing that speakers make rational decisions about redundancy and reduction and do not necessarily avoid ambiguity (Levy & Jaeger, 2007; Jaeger, 2010; Piantadosi, Tily, & Gibson, 2011). Like this previous work, we argue that ambiguity arises from a rational communication process. In our case, ambiguity arises in contexts in which the explicit costs of production are part of speakers' common ground.

Acknowledgments

This research was supported by an Andrew W. Mellon postdoctoral fellowship to the first author. The results from this paper have been presented at the 25th Annual CUNY Conference on Human Sentence Processing in New York and the 36th Penn Linguistics Colloquium in Philadelphia, both in March 2012. We thank Judith Degen, Roger Levy, and Kenny Smith for helpful discussion. We also thank research assistants Elizabeth Mazzocco, Anant Shah, Alex Djalali, and John Lee.

[3]Clark (2011) characterizes the role of salience in reduction (here, pronominalization) in a game-theoretic framework:

> It is cheaper to refer to a prominent element with a pronoun. It is correspondingly more marked (hence, costlier) to refer to a more prominent element with a description or name when a pronoun could be used. Here the speaker and hearer are presumably using some notion of salience to guide their choice. (p. 252)

References

Agresti, A. (2002). *Categorical data analysis*. Wiley.

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33(5)*, 914–930.

Arnold, J. E., & Tanenhaus, M. K. (2007). Disfluency effects in comprehension: How new information can become accessible. In E. Gibson & N. Perlmutter (Eds.), *The processing and acquisition of reference.* MIT Press.

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

Clark, R. (2011). *Meaningful games: Exploring language with game theory*. MIT Press.

Davis, W. (2010). Implicature. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2010 ed.).

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

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (p. 410-419).

Horton, W. S. (2008). A memory-based approach to common ground and audience design. In I. Kecskes (Ed.), *Intention, common ground, and the egocentric speaker-hearer* (pp. 189–222). Mouton de Gruyter.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage information density. *Cognitive Psychology*, *61*(1), 23–62.

Jäger, G. (2008). Applications of game theory in linguistics. *Language and Linguistics Compass*, *2/3*, 408–421.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman

(Eds.), *Advances in neural information processing systems.* Cambridge: MIT Press.

Lewis, D. (1969). *Convention*. Harvard University Press.

Oldenfield, R., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology*, *17(4)*, 273–381.

Piantadosi, S., Tily, H., & Gibson, E. (2011). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Schlangen, D., & Fernandez, R. (2007). Speaking through a noisy channel: Experiments on inducing clarification behaviour in human-human dialogue. In *Proceedings of Interspeech 2007 (ICSLP2007).*

Stalnaker, R. (2005). Saying and meaning, cheap talk and credibility. In A. Benz, G. Jäger, & R. van Rooij (Eds.), *Game theory and pragmatics* (p. 83-100). Palgrave Macmillan.

van Deemter, K. (2009). Utility and language generation: The case of vagueness. *Journal of Philosophical Logic*, *38(6)*, 607-632.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

There is no common ground in human-robot interaction

Geert-Jan M. Kruijff
Language Technology Lab
DFKI GmbH
Campus D3_2
Stuhlsatzenhausweg 3
D-66123 Saarbruecken, Germany
gj@dfki.de
http://www.dfki.de/~gj

Common ground is one of the key constructs in (spoken) dialogue. It embodies the notion that interlocutors build up, maintain, align their understandings of what is being talked about such that they have a "shared" or (at least) "mutual" understanding. What I understand is how you understand is what you understand. Leaving issues aside whether such a symmetry actually does exist between humans, it should be clear to anyone that it certainly does not hold between humans and robots. Robots and humans experience reality differently, they understand reality differently. There is an inherent asymmetry between them. And that creates a bit of a problem for building up shared understanding, especially since that is what we are particularly interested in, at least in the kinds of task-oriented dialogues we are often dealing with in human-robot interaction (HRI).

The talk starts by looking at what makes situated dialogue in human-robot interaction an interesting and often amusing field of research. Videos from a wide variety of projects illustrate typical problems, issues, and possibilities encountered in lab settings, as well as "out-in-the-field" (hospitals, rescue missions). From there the talk then moves deeper into the issue of common ground, looking at how it does (or where it actually does *not*) affect communication between humans and robots, and to what extent existing theories can actually really deal with the issue. The talk ends with outlining ongoing work on formulating a new approach to modeling common ground in a constructive way, and setting these efforts in practical experience with collaborative dialogue in human-robot teams.

The semantics of feedback

Harry Bunt
TiCC, Tilburg Center for Cognition and Communication
Department of Communication and Information Sciences, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, Netherlands
harry.bunt@uvt.nl

Abstract

This paper proposes a formal semantics for feedback acts in terms of updates of the information states of dialogue participants. A wide range of forms and functions of feedback is considered, including feedback about one's own processing of previous dialogue contributions ('auto-feedback') and feedback about someone else's processing ('allo-feedback'); positive and negative feedback; articulate and inarticulate feedback (having or not having a specified semantic content); feedback which is specific for a certain level of processing and feedback which is level-unspecific, and explicit feedback versus feedback that is entailed or implicated.

1 Forms and functions of feedback

Feedback is the mortar of conversation. Throughout a dialogue, the participants continuously give and elicit information about their attention, perception, understanding, and reactions to what is said by others (Allwood et al., 1993; Clark & Krych, 2004). Feedback is not always expressed explicitly through words or gestures, but may also be implicit, as in the following dialogue fragment:

(1)
1. C: Can you tell me from which platform the train to Utrecht leaves?
2. S: That's platform 5.
3. C: Thank you.

The utterance *"Thank you"* will in this situation be interpreted as implying that participant C understood S's answer, and thus as providing positive feedback by implication.

In general, the receiver of feedback obtains information about the success of his actions. In a dialogue, the receiver of a feedback message obtains information about the sender's success in processing previous contributions to the dialogue. This may tell the receiver for example that he has been understood correctly, or that the speaker is uncertain about what was meant, or has difficulty to believe something that was said. Feedback can thus relate to various levels of processing, such as hearing, understanding, and accepting something. Sometimes, a feedback message is not specific about a particular aspect of processing; for example, common forms of positive feedback such as nodding or saying *"okay"* are often ambiguous in this respect.

1.1 Auto- and allo-feedbeck

Feedback utterances most often provide information about the *speaker's* success in processing previous utterances, but they may also provide information about the speaker's beliefs about *the addressee's success* in processing. Examples are:

(2) a. A: I don't have a good connection on Thursday.
B: I said Tuesday.

b. A: Could you enhance the contrast please?
B: Is this okay?

c. A: Friday 13?
B: That's what I meant.

This kind of feedback was first distinguished by Bunt (1999) and called *'allo-feedback'*, introducing for contrast the term 'auto-feedback' to refer to feedback about the speaker's own processing. Both auto- and allo- feedback can be positive, reporting successful processing, and negative, reporting on processing that is not entirely successful.

Allo-feedback also includes *feedback elicitation*, where the speaker wants to know whether the addressee successfully processed a previous utterance. Like reportative feedback, feedback elicitation may indicate a specific level of processing, like (3c) and (3d) or may be level-unspecific. like (3a) and (3b).

(3) a. Okay?
    b. Right?
    c. Did you hear me?
    d. See what I mean?

1.2 Articulate and inarticulate feedback

A distinction among different forms of feedback concerns the specificity of the feedback. We call feedback *inarticulate* if it reports positively or negatively about the processing of (parts of) one or more previous utterances without specifying which stretch of dialogue the feedback is about (the *scope*) of the feedback), or what the result of the processing or the processing problem was. More precisely, inarticulate positive feedback reports that the processing of (parts of) one or more previous utterances was successful without specifying the scope of the feedback, or what was the result of the processing; negative inarticulate feedback reports that the processing of the utterance parts in its scope was not entirely successful, without specifying the scope or the processing problem. The examples in (4) illustrate this form of feedback.

(4) a. OK. Yes. M-hm. Aha. *(verbally expressed positive auto-feedback)*
    Nodding; smiling *(nonverbally expressed positive auto-feedback)*
    In combination: *multimodal positive auto-feedback*

    b. Excuse me? Huh? What? *(verbal negative auto-feedback)*
    Frowning; raising eye brows; head shake *(nonverbal negative auto-feedback)*
    In combination: *multimodal negative auto-feedback*

    c. Quite. Yes. (positive allo-feedback)
    Nodding *(nonverbal positive allo-feedback)*
    In combination:*multimodal positive allo-feedback*

    d. OK? All right? *(verbal negative allo-feedback)*
    Raising eye brows, looking at addressee *(nonverbal negative allo-feedback)*
    In combination: *multimodal negative allo-feedback*

Petukhova (2011) found that in the AMI corpus of multiparty dialogues, inarticulate auto-feedback is expressed only verbally in 24.2% of the cases; only nonverbally in 29.6%; and in multimodal form in 46.2%.

In contrast with inarticulate feedback, *articulate feedback* indicates the stretch of dialogue that the feedback is about, typically by repeating or paraphrasing it, and thereby also specifying a processing result. The examples in (5) illustrate this form of feedback.

(5) a. C: Which flights do you have on Friday, in the morning?
    S: To Munich, Friday the 23rd, the first flight is at 7.45. (*articulate positive auto-feedback*)

    b. Did you say Tuesday or Thursday? *(articulate negative auto-feedback)*

    c. Thursday, yes. *(articulate positive allo-feedback)*

    d. No, Tuesday. *(articulate negative allo-feedback)*

While inarticulate positive feedback is often expressed nonverbally, articulate feedback is typically expressed verbally or in multimodal form with a verbal component, since the specification of a (part of a) previous utterance and of a processing result is difficult to realize nonverbally (though an iconic or a pointing gesture can sometimes be used for that purpose).

Note that positive articulate feedback need to articulate its scope by repeating or paraphrasing the entire utterance(s) that it contains; often, only a part is repeated or paraphrased, as the examples in (6) illustrate. The paraphrase in (6a) of *"next Friday"* as *"Friday the 13th"* should be understood as positive feedback about the entire previous utterance at the level of understanding.

(6) a. B: We meet again next Friday?
    A: Friday the 13th at one-thirty.

    b. C: Can you tell me what time is the first train to the airport on Sunday?
    S: The first train on Sunday,... let me see..., the first train is at five fifty-four.

By contrast, negative feedback about part of an utterance should not be understood as negative feedback about the entire utterance, but rather as implicating positive feedback about the rest of the utterance, as the examples in (7) illustrate:

(7) a. A: Avon to Bath is four hours.
B: Four?

b. A: then go past the mill, going north,...
B: slightly northeast?

Note that the articulate/inarticulate distinction is one of (linguistic) *form*. A feedback act which is expressed in an inarticulate form does have a semantic content; the difference is that this content is provided by the utterance that the feedback is about, rather than by the feedback utterance itself.

1.3 Feedback scope

For the interpretation of feedback it is essential to know its scope. While articulate feedback explicitly indicates its scope, inarticulate feedback does not. Very often, feedback has the last utterance of the previous speaker as its scope, but not always. An analysis of the scope of feedback behaviour in two corpora, the AMI corpus[1] and a French corpus of two-party route explanation dialogue collected at the University of Toulouse[2] Petukhova et al. (2011) shows that feedback mostly (in 61% of the cases) has the immediately preceding utterance as its scope.[3] Table 2 shows the percentage of feedback occurrences with a scope of 1-10 utterances or a much larger scope (namely the entire preceding dialogue), and the distance between the feedback and its scope. We see that around 80% of the feedback cases has its scope in the preceding 1-3 utterances.

1.4 Feedback studies and statistics

Feedback has been studied empirically for its forms, functions, and contexts of occurrence, e.g. by Allwood et al. (1993), Allwood & Cerrato (2003), Clark & Krych (2004), Petukhova & Bunt (2009b), Petukhova et al. (2011), and within the conversational analysis tradition notably by Drew (1997) and Drew & Heritage (1992).

[1]See `http://www.ami-project.org`

[2]For more information see Muller & Prévot (2003).

[3]In fact this percentage is higher, since distance was measured in terms of 'functional segment's, which are smaller than utterances. See Petukhova et al. (2011) for details and for the precise definition of distance.

scope	feedback	distance	feedback
1	54.1	0	61.0
2	9.8	1	8.8
3	7.7	2	9.3
4	1.1	3	4.9
5	3.9	4	2.7
6	2.8	5	2.2
7	1.1	6	2.2
8	0.2	7	1.1
9	0.0	8	0.8
$\geq$ 10, <600	14.9	10	0.5
>600	4.4	>20	1.6

Table 1: Feedback scope and distance

Table 2 shows the frequency of occurrence of explicit feedback acts in three different corpora, the AMI corpus, the Dutch DIAMOND corpus of telephone dialogues with a help desk[4], and the OVIS corpus of Dutch human-computer telephone dialogues.[5]

	AMI	DIAMOND	OVIS
Auto-Feedback	20.5	19.1	24.1
Allo-Feedback	0.7	3.8	39.2

Table 2: Frequency (percentage of functional segments) of feedback acts in AMI, DIAMOND, and OVIS corpora.

2 Feedback as dialogue acts

2.1 Dialogue acts

Communicative feedback can be described in terms of communicative actions, performed by a speaker in order to provide information to his addressee(s) or to elicit information from him/them about the processing of previous utterances. We analyse feedback behaviour therefore within a framework constructed around communicative actions used in dialogue, called *dialogue acts*. In this framework, called Dynamic Interpretation Theory (DIT), communicative behaviour is viewed as consisting of actions that are intended to change an addressee's information state in certain ways. Such a view, commonly known as the information-state update approach to the semantics of dialogue utterances, has widely been adopted

[4]See Geertzen et al. (2004)

[5]Corpus of dialogues over the telephone with the experimental Dutch public transportation information system. See `http://www.let.rug.nl/~vannoord/OVIS`.

for the analysis of spoken and multimodal dialogue (see e.g. Larsson & Traum, 2000). The DIT framework (Bunt, 1994; 2000) has been used in the construction of a comprehensive domain-independent dialogue act taxonomy, the DIT^{++} taxonomy.[6] This taxonomy has formed the basis of the recently established ISO standard 24617-2 for dialogue act annotation (ISO 24617-2:2012); see Bunt et al. (2010; 2012).

2.2 Communicative functions and dimensions

In the DIT framework communicative behaviour is analyzed as the performance of several parallel activities, such as pursuing a certain task or activity, providing and eliciting feedback, taking turns, and editing one's contributions. Each of these types of activity is called a *dimension*; in total 10 dimensions are defined: *Task, Auto-Feedback, Allo-Feedback, Turn Management, Time Management, Contact Management, Discourse Structuring, Own Communication Management, Partner Communication Management*, and *Social Obligations Management* (see Bunt, 2009; Petukhova & Bunt, 2009a) Dialogue acts are the actions that dialogue participants use to perform these activities. A dialogue act has as its main components a *semantic content*, which specifies the entities, relations, propositions, events, actions, etc. that the dialogue act is about, and a *communicative function*, that specifies how an addressee should use the semantic content to update his information state.

A distinctive feature of the DIT^{++} taxonomy is that it consists of two parts, the 'dimension-specific' functions that can be used only for a dialogue act in a specific dimension (such as Take Turn and Turn Release in the Turn Management dimension, Stalling in the Time Management dimension, and Self-Correction in the Own Communication Management dimension), and the 'general-purpose' functions, that can be used in any dimension, such as Inform, Question, Answer, Confirm, Offer, Request, Suggest.

Figure 1 shows the taxonomy of general-purposse communicative functions, which is shared by DIT^{++} and the ISO 24617-2 standard; Figure 2 shows the DIT^{++} taxonomy of dimension-specific communicative functions, of which the ISO 24617-2 standard uses a subset.

[6]See Bunt (2009) and `http://dit.uvt.nl`.

2.3 Feedback acts

Feedback acts can be formed in two ways: (a) by combining a general-purpose function (GPF) with a semantic content that refers to the processing of previous utterances; and (b) by using a dimension-specific feedback function (FSF). GPFs can be used to form an articulate feedback act, as illustrated by the examples in (9), where we see e.g. an Auto-Feedback Set-Question in (a), an Allo-Feedback Confirm in (c), and an Auto-Feedback Inform in (d).

Both articulate and inarticulate feedback can be specific or unspecific about a level of processing; Petukhova & Bunt (2009b) show for example that inarticulate positive feedback in the form of nodding can indicate whether it is concerned with understanding or with evaluation by the speed, the number, and the amplitude of the nods. But sometimes speakers do not commit to a level of processing, in which case a level-unspecific feedback act should be used to describe the behaviour.

In DIT^{++} five levels of processing are distinguished; ordered from 'low' to 'high', these are:

(8) attention
  perception
    interpretation
      evaluation
        execution

'Evaluation' should be understood here in relation to the information-state update approach and the requirement that information states at all times remain internally consistent. For example, the recipient of an inform act with a semantic content p knows that the speaker wants him to insert the information p in his information state. Before doing this, the recipient has to check whether p is consistent with his current state; the information p is therefore buffered in the 'pending context'. If the evaluation has a positive outcome, then the recipient can move on to the stage of execution, which is the highest level of processing of an input. For this example, execution would be that the recipient moves the content from the pending context into his information state. The examples in (9) illustrate the occurrence of feedback acts relating to each of the five levels of processing.

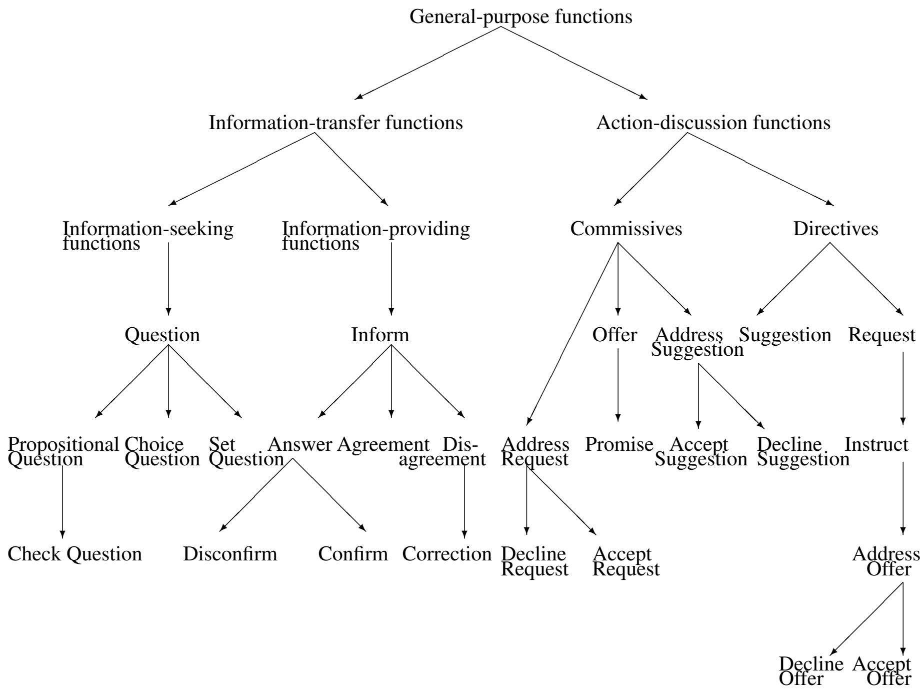


Figure 1: Dimension-speccfic communicative functions for feedback.

(9) a. Sorry, I wasn't listening. You were saying?

b. Between 11 and 1 you said?

c. A: Friday the 13th?
B: That's what I mean.

d. That's a good question.

e. A: To change the contrast first press F9.
B: Done.

The five levels of (8) have logical relationships; e.g., a message has to be perceived to some extent in order to be understood. These relations are the basis of entailments between feedback acts at different levels: a positive feedback act at one level logically entails positive feedback at lower levels, and a negative feedback act at one level entails negative feedback at higher levels. 'Positive' feedback means the utterance(s) concerned (or a dialogue act that they express) has been processed with sufficient success to not require a clarification or correction before moving on.

Moreover, the ordering of processing levels gives rise to conversational implicatures that derive from the Gricean principle of informativeness. If, for example, you did not understand well enough what was meant, then this is what you should report, rather than a perceptual problem. Therefore, positive feedback at one level implicates negative feedback at higher levels. For negative feedback it's the other way round.[7] This is summarized in Table 1.

polarity	*levels*	*relation*	*polarity*
positive	$L_i > L_j$	entailment	positive
positive	$L_i < L_j$	implicature	negative
negative	$L_i < L_j$	entailment	negative
negative	$L_i > L_j$	implicature	positive
elicitation	$L_i > L_j$	implicature	pos. allo-fb

Table 3: *Entailments and implicatures between feedback acts at different levels of processing (from Bunt, 2011b.)*

2.4 Dialogue act semantics

Bunt (2011a) formalizes communicative functions as specifications for updating an information state with a given content. This formalization has the form of a semantics for the Dialogue Act Markup Language (DiAML), defined as part of ISO standard 24617-2. In this language, a dialogue act is characterized by a sender, one or more addressees, a communicative function, a dimension, and possibly additional specifications of certainty, conditionality, or sentiment (so-called 'qualifiers'), and relations with

[7]An expression such as *"What are you saying?"* can be used to express e.g. astonishment or disbelief, rather than a perceptual problem; this is typically indicated by the use of prosody and accompanying facial expression and gestures.

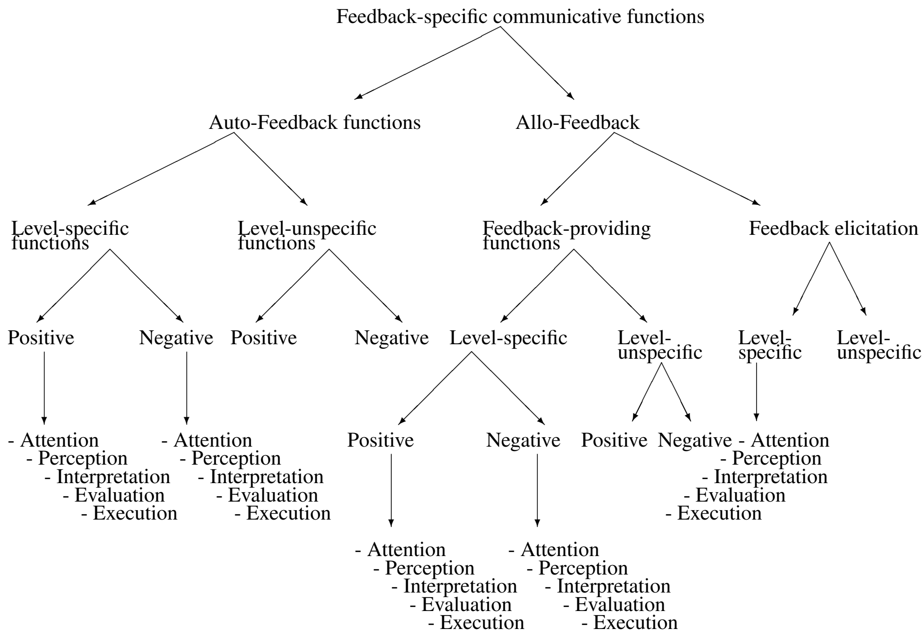


Figure 2: Dimension-speccfic communicative functions for feedback.

other dialogue acts. A specification of values for each of these parameterss gives a function that can be applied to a semantic content, resulting in an information state update operation.

3 The semantics of feedback acts

3.1 Feedback-specific communicative functions

Figure 3 shows the dimension-specific communicative functions of the DIT taxonomy for the dimensions of Auto- and Allo-Feedback. For auto-feedback there are five level-specific positive and five negative functions; likewise for allo-feedback, which has additionally five level-specific functions for feedback elicitation. In addition there are four level-unspecific communicative functions.

3.2 Semantic primitives

An analysis of the definitions[8] of the communicative functions of Fig. 1 and Fig. 2 shows that a formal description of the update effects of dialogue acts with a GPF (general-purpose communicative function) requires a number of general concepts, such as *believes that, knows value of, has goal, is able to do, is willing to do*, and that for describing the update semantics of dimension-specific communicative functions a number of dimension-specific primitives are needed. Auto- and allo-feedback acts require the following primitive predicates: **Attended, Perceived, Understood, Accepted,** and **Executed**.

[8]The definitions can be found in ISO 24617-2:2012 and on `http://dit.uvt.nl`.

3.3 Level-specific feedback acts

The semantics of level-specific feedback acts, providing information about the success of processing at level L_i, expresses that the sender of the feedback wants the addressee to know in the case of positive feedback that the utterances within its scope were successfully processed at that level; in the negative case that a processing problem occurred at that level; and in the case of feedback elicitation that the sender wants to know whether the addressee's processing was successful at that level.

The interpretation of a positive feedback act is that an addressee's information state is updated with the information that speaker wants the addressee(s) to know that the utterances in its scope were successfully processed at level L_i. This can be formalized by means of combinations of elementary update schemes in order to add two relevant beliefs to the pending context part of an addressee's information state: (1) that the speaker believes he successfully processed its content at level L_i; (2) that he wants the addressee to know that.

For example, a positive feedback act at the level of understanding, like in (6a), would be interpreted as the combination, defined in (10a), of the elementary update schemes U_{33} and U_{53} (defined in Table 5). Applied to example (6a), the update effects are that B's information about A's processing (i.e. B's pending context B'_{PC}), is extended (indicated by the symbol $=+$) to include the information that A believes he heard B say *"We meet again next Friday"*, and that A wants B to know that.

(10) a. F(AutoPerceptionPositive) $= \lambda X.\lambda Y.\lambda z. U_{33}(X, Y, z) \sqcup U_{53}(X, Y, z)$

b. B'_{PC} =+ **Bel**(B, **Want**(A, **Bel**(B, **Understood** (A, 'we meet again next friday'))))
B'_{PC} =+ **Bel**(B, **Bel**(A, **Understood**(A,'we meet again next friday')))

Table 4 lists the semantics of 5 of the 25 level-specific communicative functions of the DIT^{++} taxonomy, one for each level of processing; Table 5 shows the elementary update schemes involved. The semantics of the remaining (20) functions and update schemes can be extrapolated from these tables. For example, a positive auto-feedback act by A at the level of evaluation, addressed to B, with content c_0, updates B's pending context (B'_{PC}) using the update schemes U_{34} and U_{54} as follows:

B'_{PC} =+ **Bel**(B, **Bel**(A, **Accepted**(X, c_0))) $\sqcup$
B'_{PC} =+ **Bel**(B, **Want**(A, **Bel**(B, **Accepted** (A, c_0)))

3.4 Level-unspecific feedback acts

For determining the semantics of a feedback act which is underspecified for a level of processing, a maximally cautious approach would be to assume level-unspecific feedback to apply at the lowest level of processing, i.e. positive feedback as signalling attention without making any assumptions about signal recognition, understanding, and higher processing, and negative feedback as signalling an attention problem, and therefore also problems at all higher levels of processing. This does not seem realistic, however; level-unspecific positive feedback signals like *"yes"*, *"okay"*, and nodding typically signal more than just paying attention, and negative signals do not just signal a problem at the level of attention, but rather at a higher level. We propose to determine the levels of processing covered by level-unspecific feedback acts empirically.

To this end, we analyzed the feedback level interpretations in data obtained in an annotation experiment, originally performed in order to assess inter-annotator agreement among naive annotators using the DIT^{++} annotation scheme (see Geertzen et al., 2007). The experiment showed that annotators often found it difficult to choose a level of processing when annotating flevel-unspecific eedback acts. This explains why agreement scores were found for auto- and allo-feedback of .36 and .33, respectively, which are much lower than those for other dimensions (average .61). This motivated the designers of the ISO 24617-2 annotation scheme to collapse the level-specific feedback functions of DIT^{++} into the level-unspecific communicative functions Auto-Positive, Auto-Negative, Allo-Positive, Allo-Negative, and Feedback Elicitation (which were subsequently also added to the DIT^{++} taxonomy).

We analyzed the annotations produced in this experiment for the number of times annotators assigned a particular level to a feedback act of which the level was not clearly expressed in linguistic and/or nonverbal features of the behaviour, and calculated the number of times each level was chosen in those cases where not all four annotators agreed. The results are shown in Table 4 for human-human dialogues from the Map Task corpus and for human-computer dialogues from the OVIS corpus.

The table shows that level-unspecific feedback is almost never interpreted as applying at the level of attention. For the rest, the results are very different. In the human-human condition positive auto- and allo-feedback are both interpreted mostly as applying to evaluation or execution, whereas in the human-computer dialogues most feedback acts concerned perception or understanding. The latter result is directly related to the deficiencies in automatic speech recognition, and to some degree also to the machine's limited understanding of the user.

Since the interpretation of level-unspecific feedback acts depends on the setting in which the dialogue occurs, we propose to introduce a predicate **SuccessProcessing** that represents successful processing, whose interpretation depends on the dialogue setting. For human-human dialogue (the MT

F(AutoAttentionPositive)	=	$\lambda X.\lambda Y.\lambda z.U_{31}(X,Y,z) \sqcup U_{51}((X,Y,z))$
F(AlloPerceptionNegative)	=	$\lambda X.\lambda Y.\lambda z.U_{37}((X,Y,z)) \sqcup U_{57}((X,Y,z))$
F(AutoInterpretationPositive)	=	$\lambda X.\lambda Y.\lambda z.U_{33}(X,Y,z) \sqcup U_{53}((X,Y,z))$
F(AutoEvaluationPositive)	=	$\lambda X.\lambda Y.\lambda z.U_{34}(X,Y,z) \sqcup U_{54}((X,Y,z))$
F(ExecutionElicitation)	=	$\lambda X.\lambda Y.\lambda z.U_{75}(X,Y,z)$
F(AutoPositive)	=	$\lambda X.\lambda Y.\lambda D_i.\lambda z.U_{39}(X,Y,z) \sqcup U_{59}(X,Y,z)$

Table 4: Semantics of feedback functions (selection)

$U_{31}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Want**(X, **Bel**(Y, **Attended**(X,z))))
$U_{33}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Want**(X, **Bel**(Y, **Understood**(X,z))))
$U_{34}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Want**(X, **Bel**(Y, **Accepted**(X,z))))
$U_{37}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Want**(X, **Bel**(Y, **Perception-Problem**(Y,z))))
$U_{39}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Want**(X, **Bel**(Y, **SuccessProcessing**(X,z))))
$U_{51}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Bel**(X, **Attended**(X,z))))
$U_{53}(X,Y,D_i,z)$	Y'_{PC} =+ **Bel**(Y, **Bel**(X, **Understood**(X,z))))
$U_{54}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Bel**(X, **Accepted**(X,z))))
$U_{57}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Bel**(X, **Perception-Problem**(X,z))))
$U_{59}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Bel**(X, **SuccessProcessing**(X,z))))
$U_{75}(X,Y,z)$:	Y'_{PC} =+ **Bel**(Y, **Want**(X, **Know-if**(X, **Execution-Problem**(Y,z))))

Table 5: Elementary update schemes used in the semantics of auto- and allo-feedback functions.

condition), according to Table 6 this predicate can be interpreted as representing successful processing at the level of understanding or higher, i.e., as signalling successful understanding and possibly also successful 'higher' processing. Negative feedback would be interpreted as complementary to positive feedback.

So a positive level-unspecific feedback act, with the communicative function AutoPositive, like the one contributed by B in (11a), would (according to Table 4) be interpreted by the combination of elementary update schemes U_{39} and U_{59}, defined in Table 5.

(11) a. A: I said five buttons max.
B: Okay.

b. F(AutoPositive) = $\lambda X.\lambda Y.\lambda z.U_{59}(X,Y,z) \sqcup U_{39}(X,Y,z)$

c. B'_{PC} =+ **Bel**(B, **Want**(A, **Bel**(B, **SuccessProcessing** (A, 'I said five buttons max'))))
B'_{PC} =+ **Bel**(B, **Bel**(A, **SuccessProcessing**(A, 'I said five buttons max')))

Interpreting the predicate **SuccessProcessing** in certain conditions as "well understood and possibly also accepted and executed successfully" may seem to mean that level-unspecific feedback is in fact interpreted as level-*specific* feedback at the level of understanding, but there is a subtle difference in implicatures: level-specific positive feedback at the level of understanding implicates negative feedback at the levels of evaluation and execution, but in the case of level-unspecific feedback these implicatures do not arise. Similarly for implicated negative feedback.

3.5 Entailed and implicated feedback

Feedback may be entailed or implicated by non-feedback acts. Example (1) illustrated the occurrence of *implicated* positive feedback, which is at the highest level of processing (the answer that the thanking applies to is not just *understood*, but also *accepted* and *adopted*). Negative feedback may be implicated e.g. when the speaker jumps abruptly to a new topic, which may carry the suggestion that the previous topic was closed in an unsatisfactory manner; in such a case it is not evident at which level of processing a problem occurred.

Positive feedback is *entailed* by all responsive dialogue acts such as answers, confirms and disconfirms; acceptance or rejection of offers, suggestions, or requests; return greetings, accept apologies, and several others.

(12) a. A: So, um, how many buttons do you suggest?
B: I said five max.

feedback	*auto-positive*		*auto-negative*		*allo-positive*		*allo-negative*	
level	MT	OV	MT	OV	MT	OV	fMT	OV
attention	3	0	0	0	0	0	0	0
perception	0	28	0	6	1	3	0	9
interpretation	4	20	0	6	2	0	0	14
evaluation	32	0	1	0	8	0	0	6
execution	34	0	1	8	12	0	0	0

Table 6: Interpretation levels (in percentages) in Map Task (MT) dialogues and OVIS (OV) dialogues)

b. A: Pete, could you start the presentation?
   B: Sure.

c. A: Sorry, we have no information about that.
   B: No problem.

In all these cases it can be argued that the responsive dialogue act is only possible if the 'antecedent' dialogue act was sufficiently well understood, was accepted, and was 'executed' successfully. This illustrates that entailed feedback is in general at the highest level of processing, that of execution, and therefore at all levels.

In sum, implied positive feedback, wether entailed or implicated, is positive at all levels of processing. Implied negative feedback is virtually never about failed attention or perception, but rather about understanding, evaluation, or execution.

4 Applications

The study of the forms, functions, and semantics of feedback has both theoretical and practical applications.

Theoretically, a good understanding of feedback is indispensable for a good understanding of language in interaction, and has been studied in relation to natural language understanding e.g. by Ginzburg (1994), Ginzburg & Cooper (2004), Purver et al., (2001). Feedback plays a crucial role in processes of grounding (the establishment of common ground among dialogue participants), and has as such been studied e.g. by Traum (1994), Clark (1996) and Bunt et al. (2007). The semantics of feedback plays a role in some of these studies, but often not in an explicit and certainly not in a complete way; for example, allo-feedback has not been considered in any of these studies with the exception of Bunt (1999).

Sophisticated interactive automatic systems should be able to understand and to generate appropriate forms of feedback at appropriate points in the interaction. This application of models of feedback has been investigated e.g. by Van Dam (2006), for the design of graphical user interfaces; for designing the PARADIME dialogue manager of the IMIX information extraction system (Keizer et al., 2011), for the design of the multimodal DENK dialogue system (Ahn et al., 1995) and for the GoDiS dialogue system by Larsson et al. (2000).

5 Conclusions

In this paper we have shown that a formal and computational semantics in terms of information state updates can be given for a wide range of forms and functions of feedback, including auto- and allo-feedback (including feedback elicitation), which both can be positive or negative, articulate and inarticulate, specific for a particular level of processing or level-unspecific, and entailed or implicated.

For feedback acts which are unspecific regarding a level of processing, we proposed to use an empirically determined level of success, which appears to be different for human-computer dialogue than for natural human dialogue.

References

Ahn, R., R.-J. Beun, V. Borghuis, H. Bunt and C. van Overveld (1995) The DENK architecture: A fundamental approach to user interfaces. *Artificial Intelligence Review* 8(3): 431-435.

Allwood, J., J.Nivre and E. Ahlsén (1993) On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics* 9, 1–26.

Allwood, J.,and L. Cerrato (2003) A study of gestural feedback expressions. *Proceedings of the First Nordic Symposium on Multimodal Communication*, pp. 7–22.

Bunt, H. (1994). Context and Dialogue Control. *Think Quarterly* 3 (1), 19–31.

Bunt, H. (1999). Dynamic interpretation and dialogue theory. In M. Taylor, D. Bouwhuis and F. Néel (eds) *The Structure of Multimodal Dialogue, Vol. 2.* Benjamins, Amsterdam, pp. 139–166.

Bunt, H. (2000). Dialogue pragmatics and context specification. In H. Bunt and W. Black (Eds.), *Abduction, Belief and Context in Dialogue.* Amsterdam: John Benjamins, pp. 81–150.

Bunt, H. (2009). The DIT^{++} taxonomy for functional dialogue markup. In *Proc. AMAAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts"*, pp. 13–25.

Bunt, H. (2011a). The semantics of dialogue acts. In *Proceedings 9th International Conference on Computational Semantics IWCS 2011*, Oxford, pp. 1–24.

Bunt, H. (2011b). Multifunctionality in dialogue. *Computer, Speech and Language* (25), 225 – 245.

Bunt, H., R. Morante and S. Keizer (2007) An empirically based computational model of grounding in dialogue. In *Proceedings 8th SIGDIAL Workshop on Discourse and Dialogue,* Antwerp, pp. 283-290.

Bunt H, Alexandersson J, Carletta J, Choe JW, Fang A, Hasida K, Lee K, Petukhova V, Popescu-Belis A, Romary L, Soria C, Traum D (2010) Towards an ISO standard for dialogue act annotation. In: *Proceedings 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

Bunt H, Alexandersson J, Choe JW, Fang A, Hasida K, Petukhova V, Popescu-Belis A, andTraum D (2012) A semantically-based standard for dialogue annotation. In: *Proceedings 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul.

Clark, H. (2006). *Using Language*. Cambridge University Press.

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

Dam, H. van (2006) Dialogue acts for GUIs. Ph.D. Thesis, Eindhoven University of Technology.

Drew, P. (1997) 'Open' class repair initiators in response to sequential sources of troubles in conversation, *Journal of Pragmatics* 28: 69-101.

Drew, P. and J. Heritage (eds.) (1992) *Interaction in institutional settings,* Cambridge University Press.

Geertzen, J., Girard, Y., and Morante, R. (2004) The DIAMOND Project. in *8th Workshop on the Semantics and Pragmatics of Dialogue*, Barcelona.

Geertzen, J., Petukhova, V., and Bunt, H. (2007) *A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification*. In *Proceedings of SIGdial 2007 Workshop on Discourse and Dialogue*, Antwerp, pp. 140-149.

Ginzburg, J. (1994). An update semantics for dialogue. In H. Bunt (ed.) *Proceedings of the 1st International Workshop on Computational Semantics (IWCS-1)*, Tilburg, pp. 111-120.

Ginzburg, J. and R. Cooper (2004). Clarification, ellipsis, and the nature of contextual update. *Linguistics and Philosophy* 27(3): 297-366.

ISO (2012). *ISO 24617-2: Semantic annotation framework Part 2: Dialogue acts.* ISO, Geneva.

Keizer, S., H. Bunr and V. Petukhova (2011). Multidimensional dialogue management. In A. van den Bosch and G. Bouma (2011). *Interactive Multimodal Question-Answering.* Berlin: Springer, pp. 57-86.

Larsson, S., P. Ljunglöf, R. Cooper, E. Engdahl and S. Ericsson (2000). GoDiS - an accommodating dialogue system. In *Proceedings ANLP/NAACL-2000 Workshop on Coversational Systems*, pp. 7-10.

Larsson, S. and D. Traum (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering* 6: 323-340.

Muller, P. and L. Prévot (2003). An empirical study of acknowledgement structures. In *Proceedings of Diabrück, the 7th Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken.

Petukhova, V. and H.Bunt (2009a) The independence of dimensions in multidimensional dialogue act annotation. In *Proceedings NAACL 2009 Conference* , Boulder, pp. 197-200.

Petukhova, V. and H.Bunt (2009) Grounding by nodding. In *Proceedings of GESPIN, Conference on Gestures and Speech in Interaction*, Poznán.

Petukhova, V., L. Prévot and H.Bunt (2011) Multi-level discourse relations between dialogue units. In *Proceedings 6th International Workshop on Interoperable Semantic Annottion (ISA-7)*, Oxford, pp. 18–28.

Purver, M. (2004). *The Theory and Use of Clarification in Dialogue.* Ph.D. Thesis, University of California, Berkeley, USA.

Purver, M., J. Ginzburg and P. Healy (2003). *On the Means for Clarification in Dialogue.* In R. Smith & J. van Kuppeveldt (eds) *Current and New Directions in Discourse and Dialogue*, Kluwer, Dordrecht.

Traum, D. and S. Larsson (2003). The information state approach to dialogue management. In R. Smith & J. van Kuppeveldt (eds) *Current and New Directions in Discourse and Dialogue*, Kluwer, Dordrecht.

Traum (1998). A computational theory of grounding in natural language conversation. PhD Thesis, University of Rochester.

Recovering from Non-Understanding Errors in a Conversational Dialogue System

Matthew Henderson
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge, CB2 1PZ
mh521@cam.ac.uk

Colin Matheson
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
colin@inf.ed.ac.uk

Jon Oberlander
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
jon@inf.ed.ac.uk

Abstract

Spoken dialogue systems can encounter different types of errors, including non-understanding errors where the system recognises that user has spoken, but does not understand the utterance. Strategies for dealing with this kind of error have been proposed and tested in the context of goal-driven dialogue systems, for example by Bohus with a system which helps reserve conference rooms (Bohus and Rudnicky, 2005). However there has been little work on possible strategies in more conversational settings where the dialogue has more open-ended intentions. This paper looks at recovery from non-understanding errors in the context of a robot tourguide, and tests the strategies in a user trial. The results suggest that it is beneficial for user enjoyment to use strategies which attempt to move the dialogue on, rather than getting caught up in the error by asking users to repeat themselves.

1 Introduction

The handling of understanding errors is an important consideration in the design of a spoken dialogue system. Many dialogues take place in difficult conditions, with spontaneous speech, large vocabularies, varied user populations and uncertain line quality (Bohus and Rudnicky, 2005). These conditions make understanding errors very likely during the course of a dialogue.

There are two types of understanding error which a spoken dialogue system may encounter: *non-understandings* and *misunderstandings*. A non-understanding is where the system fails to extract a valid semantic representation of what the user said. A misunderstanding is where a valid representation is extracted which happens to be incorrect. While detecting misunderstandings requires some thought, non-understanding errors are immediately apparent to the system due to the failure of the natural language understanding component.

This paper looks at strategies for dealing with non-understanding errors in the context of conversational spoken dialogue systems, as opposed to slot-filling or more generally goal-driven approaches. In such goal-driven systems, the user and the system typically work together to accomplish a specific task, for example booking a flight, finding a restaurant or reserving a conference room. This normally involves the system obtaining some information from the user (or filling a list of slots with their values), checking a database, and then completing the task. In a more conversational dialogue system the only real task is to take part in an interaction which is interesting and enjoyable for the user, although in work related to ILEX (Mellish et al., 1998) the system may have the loose goal of communicating prioritised pieces of information, and the research reported here is in this tradition. There may not be a definitive distinction to be drawn between what we have termed ‘conversational’ and ‘goal-driven’ systems, apart from pointing to the typical need that the latter have to fill slots with information elicited from the user, while there is no such target in the former.

The INDIGO project (Vogiatzis et al., 2008), (Konstantopoulos et al., 2009) followed the ILEX notion of opportunistic language generation, adapting the approach to spoken interactions with a robot museum guide. A later version of the guide was tested with an initial 'fake' strategy (see below) for avoiding repetitions of the standard '*Could you please repeat that*' form for dealing with non-understandings. This unreported pilot work is updated here and extended to include a set of non-understanding error recovery strategies which aim to improve user enjoyment of conversational dialogues with a robot tourguide. The strategies are tested in a user trial which is designed to elicit answers to the following questions:

- Can user satisfaction be increased by using smart strategies to deal with non-understanding errors in a conversational dialogue system?
- How does the use of such strategies affect the user's perception of the dialogue and the dialogue system?
- How do the strategies compare to each other in terms of user satisfaction, and in particular is it important to employ a variety of strategies?

2 The Tourguide Dialogue System

The Tourguide Dialogue system was built in order to investigate non-understanding error recovery strategies in a conversational domain. The chosen application is that of acting as a tourguide in an exhibition. The dialogues consist of the system describing an item, and then taking questions from the user. Specifically, the system talks about 3 items which can be found in the Informatics Forum at the University of Edinburgh.

During the course of the tourguide dialogues, the point where most errors are anticipated is when the system asks 'Do you have any questions about this?' This obviously constitutes an extremely open question, and the lack of constraints on the user's input results in a high probability of a non-understanding occurring. At all the other points where the system elicits input from the user it has full initiative and can supply the speech recognition module with a set of highly constrained expectations, whereas in the situation above (although the system does attempt to predict the input), the range if possibilities is very large. This is thus a good context in which to investigate strategies for dealing with non-understanding errors. The system is designed to be programmed with a library of error recovery strategies. For a list of the strategies implemented, see Table 1.

Table 1: Summary of Strategies

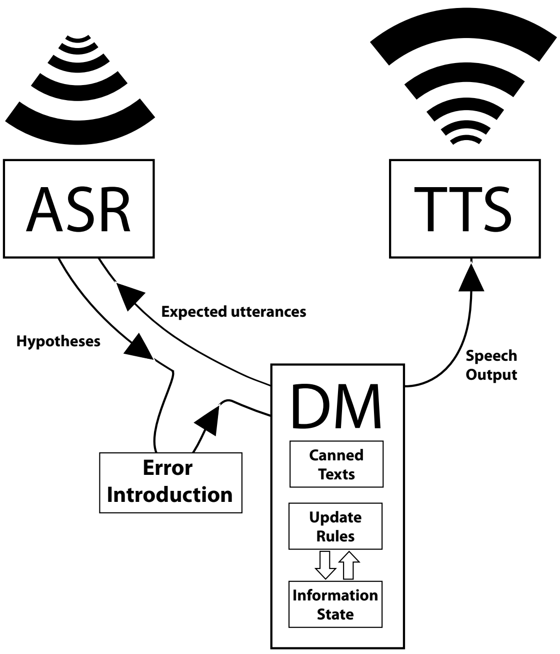
Name	Description
Subsume	Ask if the user would like to hear more information about the item E.g. : "Would you like to hear a little more about Paolozzi and his sculptures?"
Subsume Split	Ask if the user is more interested in hearing about aspect A or aspect B of the item E.g. : "Well, are you more interested in finding out more about Paolozzi himself, or his sculptures?"
Fake	Fake having forgotten to say something of interest about the item E.g. : "I meant to add; one of Paolozzi's most famous works can be found here in Edinburgh. At the top of Leith Walk, there are sculptures of human body parts, including a giant foot, by Paolozzi."
Please Repeat	Ask the user to repeat their question E.g. : "Please could you repeat that? Just say 'no' if you have no more questions."

The dialogue manager is implemented in Prolog, using the Trindikit framework. It is a handcrafted dialogue system, which uses the information state model to hold the system's beliefs and a set of update rules which define the system's actions. See Larsson and Traum (2000) for a summary of information state and dialogue management using the framework.

For speech input and output, modules developed by Acapela Group[1] for the INDIGO project were used. As mentioned, at each point in the dialogue at which user input is expected, a handcrafted list of possible user utterances is sent to the speech recogniser. For example, at a point

[1] http://www.acapela-group.com

Figure 1: Overview of the Tourguide Dialogue System



where the system asks for questions on a particular exhibit, the speech recogniser is supplied with a list of questions which were predicted by the system designer. Language generation uses only simple templates, which are sent to the Acapela text-to-speech component.

In order to ensure that non-understanding errors occur at a consistent and non-negligible rate, a component is introduced into the system between the speech recogniser and the dialogue manager which serves to introduce errors at a predetermined rate. If a real non-understanding hasn't happened after 3 questions, an error is introduced by throwing away the speech recognition result and simulating a null user act. An overview of the structure of the system is presented in Figure 1.

3 Non Understanding Error Strategies

3.1 Motivating the Strategies

There have been studies using human communication to investigate how human agents deal with non-understanding errors, in the hope that this can be applied to spoken dialogue systems (Zollo, 1999; Skantze, 2003; Koulouri and Lauria, 2009). Wizard of Oz methods allowed analysis of dialogues between human users linked by computer systems. To emulate a real spoken dialogue system, the Wizard sees the output of a speech recogniser, and the user either listens to the output from a speech synthesiser or a vocoder. All three of these studies focused on relatively restricted goal-driven dialogues, where the user and system had to work together to accomplish a task with a clear target outcome.

In these experiments, the wizards and users were always naive participants, and the responses of the wizards were not limited in any way (with the exception of some of the conditions in Koulouri and Lauria (2009)). This allowed the experimenters to analyse how a human might try to deal with speech recognition errors when trying to conduct a dialogue.

A common theme in all three studies was the importance of using error recovery strategies which help the dialogue to progress. It was found that wizards will often ask task-related questions, the answer to which subsumes the information missed by the non-understanding. The example question from Skantze (2003) below illustrates this:

wizard
Do you see a wooden house in front of you?
user
Yes crossing address now. (*Actually:* I'm passing the wooden house now.)
wizard
Can you see a restaurant sign?

Here the Wizard asks a follow-up question which is related, in that its answer implies the information they just missed. Skantze found that this strategy not only improved the understanding of the following utterances, but also resulted in higher user perception of task success.

Other wizard of Oz studies have looked specifically at evaluating error recovery strategies (Schlangen and Fernández, 2006; Rieser et al., 2005), and Bohus implemented a variety of non-understanding error recovery strategies in a real dialogue system (Bohus, 2007); relevant findings are summarised in Bohus and Rudnicky (2005). Again this study focuses on a goal-driven dialogue system, specifically a system which helps users book conference rooms. In the current context, one of the most interesting strategies implemented was called *MoveOn*, where the system would continue by asking a new question when faced with a non-understanding. An example is:

> Sorry, I didn't catch that. One choice would be Wean Hall 7220. This room can accommodate 20 people and has a whiteboard and a projector. Would you like a reservation for this room?

This strategy performed well with respect to recovery rate, i.e. how often the following user response was correctly understood. Bohus and Rudnicky explained its success by comparing it to other strategies, which would generally ask the user to repeat themselves, or rephrase their answer. In those cases it is unlikely that the system will be able understand the user's intention as it did not understand the input the first time. This process is prone to turning into a spiral of errors, with the user getting more and more frustrated. Frustration can affect the user's voice, in turn adversely affecting the Automatic Speech Recognition. On the other hand, with *MoveOn*, the system abandons the current question and tries a new line of attack.

The *MoveOn* strategy is related to the recommendations of Zollo, Skantze and Koulouri, and it seems from these studies that the idea of moving on, and asking a new question can be very effective. However it is not entirely clear how this strategy can be adapted to use in a conversational dialogue system. It is to this question that we now turn.

3.2 The Strategies

In a conversational dialogue system, there is as noted above no real goal in the sense of information to be elicited and acted upon, so it is not clear what constitutes a 'task-related' question in the sense used in the above studies. Indeed, in the Tourguide dialogue system, it is not usually the robot which is asking questions of the user but the other way around.

The general aim is thus to progress the dialogue smoothly when the user has just asked a question about an item in the exhibit which the system hasn't been able to understand. The first strategy which attempts to do this is called *Subsume* (see Table 1 for a summary of all the strategies, with examples). The Subsume strategy asks if the user is interested in finding out more about the item, it then waits for a response – any response – and then proceeds to output a short text about the item. The text is designed to incorporate answers to a lot of the possible questions which the user may have asked. The strategy tries to broaden the user's goal from obtaining a specific piece of information to just hearing some general interesting information about the piece.

The second strategy is *Subsume Split*, which is similar to Subsume but gives the user a choice of what subsuming information they prefer. The questions for every item in an exhibit should broadly be able to be split into two categories. For example, for an artefact like a sculpture, these could be (a) questions about the artefact's creator and (b) questions about the artefact and other examples of the creator's work. In giving the user a binary choice the hope is that the information subsequently presented will be of more interest, and more closely related to their original question. Whereas Subsume did not rely on the next utterance being understood, Subsume Split requires the speech recogniser to distinguish between two possible answers. This is of course back in line with standard system-initiative approaches in which speech errors are much less of a problem as speech recognition generally works well in constrained contexts.

The last strategy implemented is called *Fake*, an approach which was investigated in the non-published pilot study of the INDIGO project mentioned above. The idea is for the system to pretend it has forgotten to include a piece of information. Although it is clearly unrealistic for a robot to be forgetful, it is hoped that the piece of information which the system provides is sufficiently interesting to justify saying it instead of answering the user's actual question. In the pilot study, the system returned to the original dialogue state after the diversion, asking Now, did you have a question?, and attempted to combine the speech recogniser scores from the two user turns to gain a higher confidence hypothesis. The current version implements the same dialogue moves but does not combine the recognition scores.

As well as these strategies, the system has a default *Please Repeat* strategy which is used as a baseline. This is used as a backup strategy whenever the other strategies are exhausted in a particular dialogue.

An excerpt from a typical dialogue follows, where an error is introduced and the system uses

Table 2: Questionnaire

Communication	Agent	Attitude
The system understood what I said	The system is intelligent	I enjoyed talking to the system
My conversation with the system flowed smoothly	The system was helpful	I felt confused when talking to the system*
It was clear what was happening when the system did not understand me	My conversation with the system was interesting	I felt frustrated when talking to the system*

For further analysis, responses to questions marked with an asterisk are converted from $(1, 2, 3, 4, 5)$ to $(5, 4, 3, 2, 1)$ so that higher numbers correspond to higher user satisfaction as with the other questions.

---

the SubsumeSplit strategy.

Excerpt from Typical Dialogue

System: This is a sculpture by Eduardo Paolozzi, a sculptor from Edinburgh. Paolozzi was interested in science, and our quest to understand more about the Universe and ourselves. The themes of his work relate to the research done here in Edinburgh. This piece is different in style to usual Paolozzi sculptures in its simplicity and smoothness. Do you have any questions?

User: Umm. Where is he from?

System: Paolozzi was born in Leith, in north Edinburgh. Do you have any other questions?

User: Why did he chose this shape?

System: I'm not sure why the sculpture is this shape. You'd have to ask Paolozzi himself! Do you have any more questions?

User: What does it represent?

-**Non-understanding error introduced**-

System: Well, are you more interested in Paolozzi himself, or his sculptures?

User: Eh, his sculptures.

System: Paolozzi's sculpture work was inspired by some of the great minds in the history of science. He invented a method of casting bronze sculptures, analogous to the process of creating collages. His early collage work in Paris pre-empted Pop Art. There are 4 of his sculptures in the Informatics Forum, and his work can be seen throughout the cities of Britain. Do you have any more questions?

4 Experimental Setup

The experiment consists of running a user trial with the system in 3 different modes. In the first mode (`mixed`) the system uses all the strategies shown in Table 1. In the second (`single`) the system uses a single strategy, either Subsume Split, or Fake. And in the last (`pleaserepeat`) it has no strategies except the default Please Repeat.

Participants were sat in front of a laptop running the Tourguide Dialogue System and asked to conduct a 10 to 15 minute long dialogue. The laptop screen displays the text as it is synthesised by the system, and also a list of example responses at each stage in the dialogue. Other than the length requirement, the users were not given any particular tasks to achieve in order to approximate a natural interaction with a conversational system. The participants were shown pictures of the three items which the system can talk about, and were told to ask the questions they believe they might ask if they were actually at the exhibition with a robot.

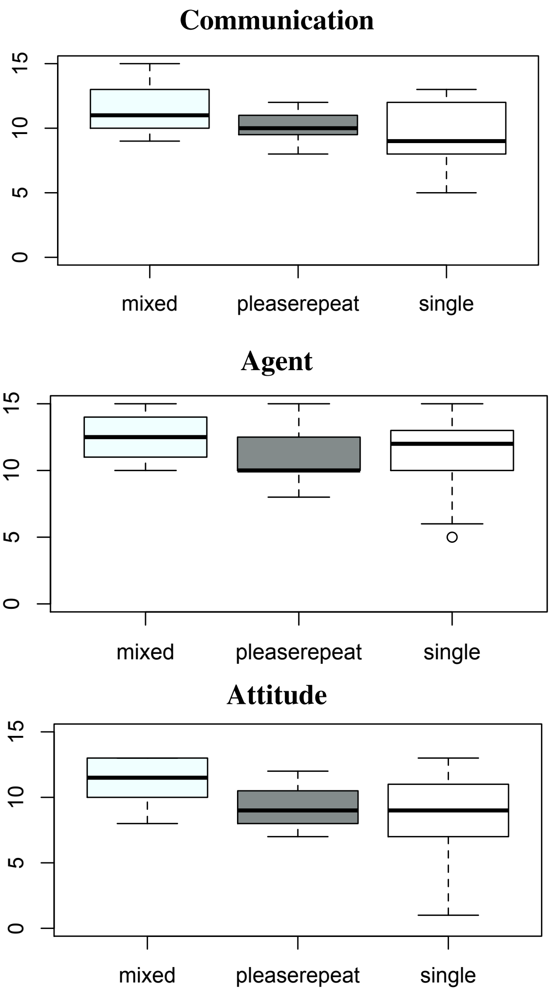
As mentioned, the system is configured to introduce a non-understanding error at every third question asked, as long as a real error did not occur in the previous three turns. The error rate is thus relatively consistent across the dialogues. The misunderstanding error rate due to incorrect speech recognition on user questions was 18%, this did not change significantly between conditions.

At the end of the interaction, participants were asked to fill in a questionnaire which includes a series of statements with which the participant must specify their level of agreement on a scale of 1 to 5. These statements are listed in Table 2, and are designed to measure the user's satisfaction along multiple dimensions. These questions serve to quantify the quality of the dialogue from the user's perspective better than an objective score such as dialogue length could estimate.

5 Results

Data from 58 participants in total was gathered, 14 in the `mixed` condition, 29 in the `single` condition (14 with Fake strategy and 15 with Subsume Split) and 15 in `pleaserepeat`. The questions on the questionnaire are grouped into three collections as shown in Table 2. The col-

Figure 2: Breakdown of Questionnaire Results



lections correspond respectively to the quality of the *Communication*, the user's perception of the system as an *Agent*, and the user's *Attitude* towards the dialogue. Within the three collections, the question answers are found to be highly correlated. The individual scores of each question in a collection are combined by simply adding them together, giving a collection score between 0 and 15.

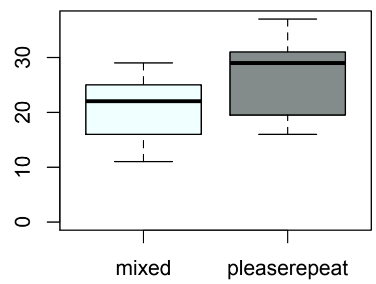
Figure 2 shows the results for each collection in each of the 4 conditions as box-whisker plots. In this paper, values more than $3/2$ times the interquartile range lower than the first quartile are treated as outliers, as are values that are analogously higher than the third quartile.

6 Analysis of Results

6.1 Analysis of Questionnaire Data

Kruskal-Wallis tests are used to test the hypothesis that the boxplots shown in Figure 2 represent distinct distributions, i.e. that there is some difference in the distribution of a collection score between the groups. These tests suggest further investigation into Communication and Attitude (with the probability of the null hypothesis being less than 0.02), but not into the Agent scores.

Figure 3: Number of Questions Asked



Pairwise Mann-Whitney tests in the Communication and Attitude data are performed to test whether the differences between the pairs of groups are significant. Bonferroni correction is used to account for the fact that there are 3 comparisons for each collection, so a threshold of 0.015 $(< 0.05/3)$ on the p-value is chosen. From this analysis, the following comparisons are found to be significant:

- **Communication**
  `mixed` > `pleaserepeat`;
  `mixed` > `single`.
- **Attitude**
  `mixed` > `single`.

These results imply that the quality of the Communication of a dialogue (recall a combination of the flow of conversation, clarity of the system's actions and how well it seems the system understands the user) is significantly improved by using a mixture of error recovery strategies as against a single strategy, as well as against the baseline please repeat. Variety in the dialogue may give the user an impression of a richer dialogue. Figure 3 shows how the number of questions asked is much higher in the `pleaserepeat` condition than in `mixed`.

This is because the mixture of strategies allow the system to do more of the talking, and to answer many of the user's questions before they are asked. Less user questions means less possibility for error, and thus better dialogues. The strategies are exploiting the fact that users don't mind being provided more information than they originally asked for.

As mentioned, the Attitude measure is a combination of user enjoyment and lack of confusion and frustration. This is found to be significantly better in the `mixed` condition than in the `single` condition, but the comparison between `mixed` and `pleaserepeat` is not statistically significant ($p = 0.05$).

Note that the `mixed` condition is at an advantage relative to the `single` condition because it will take longer before the system resorts to the Please Repeat strategy. Therefore in the comparisons we must bear in mind that there are on average more Please Repeats being issued in the `single` condition.

6.2 Discussion

The strategies effectively use errors as an opportunity to tell the user something which it believes could be of interest. In a more complex system, the information provided could be tailored using a user model, as in the approach noted in the Introduction (Mellish et al., 1998). It is worth noting that if a system can opportunistically exploit errors to actively improve user experience, it could weaken the typical inverse correlation between user satisfaction and non-understanding rate, or at least, the rate of repetition-requests (Walker et al., 2000). Demonstrating this remains a matter for future work, however, since the current study specifically maintained a constant non-understanding rate across conditions, rather than treating it as an independent variable.

Lastly, it is interesting to investigate some of the correlations between the individual questionnaire answers using Pearson's correlation tests. The users' enjoyment is not correlated with how clearly they understand what the system is doing when an error occurs. This implies that it is not necessarily important for the user to understand what motivates the system's dialogue turns for them to enjoy the interaction. This appears to contradict the findings of Hockey et al. (2003) among others, which show that making the system 'visible' to the user increases the level of task success. The suggestion is therefore that the latter finding only applies in goal-driven dialogue systems, and so although the user must have some idea of what is motivating the system, it is not necessarily as important in more conversational settings.

7 Conclusions

In summary, this study has provided evidence that these new strategies, which use the idea of moving the dialogue on when the system has little or no input from the user, can have a positive effect on overall user satisfaction. It is shown that the benefit of such strategies is in using them as a strategy, and giving a conversational dialogue system a variety of error handling techniques.

Use of all of the strategies was significantly beneficial for the dialogues in the three dimensions measured in the questionnaire. Therefore, when designing a conversational dialogue system, it is worthwhile putting thought into the design of error recovery strategies which are more complex than asking the user to repeat or rephrase themselves. It is particularly beneficial to ensure that there is a variety of strategies available to the system, both to increase the variation in the dialogue and to make the individual strategies more effective. This has been confirmed experimentally in the goal-driven domains (see Section 3.1), and this paper provides initial supporting evidence in conversational, less goal-directed applications.

8 Future Work

A number of potential further investigations are possible:

- Presumably user enjoyment in a conversational dialogue system tends to degrade as error rates increase (Walker et al., 2000). It would be interesting to compare how quickly this degradation occurs when different error recovery strategies are employed. It is possible that strategies such as those presented here would help to maintain a minimal level of enjoyment longer.
- At one end of the spectrum, some goal-driven dialogue systems can be associated with a single objective metric of task success, independent of user impressions. Towards the other end of that spectrum, conversational systems like museum tour guides should allow different visitors to pursue distinct tasks, or single visitors to shift from one task to the other, and even interleave them. In such cases, more work is needed to identify the varying criteria for success for any given user.
- In this study the `mixed` condition chooses

strategies at random. It might be useful to investigate whether there exists a better-than-random policy. Bohus et al. have looked at this question in goal-driven applications (Bohus and Rudnicky, 2005; Bohus et al., 2006).

- More strategies could be investigated, possibly ones which exploit a user model to select pieces of information to impart. The current strategies use text which is the same for all users, whereas the use of a full language generation system producing dynamic texts would not only allow for tailoring to the user but also cause the strategies to be used more than once in a given part of the dialogue.

Acknowledgements

The work reported here was supported by both the INDIGO (IST-045388) and Help4Mood (ICT-248765) projects. We are grateful to Vasilis Karaiskos for assistance in completing the evaluations.

References

Dan Bohus and Alex Rudnicky. 2005. Sorry, I didn't catch that! An investigation of non-understanding errors and recovery strategies. In *6th SIGdial Workshop on Discourse and Dialogue*.

Dan Bohus, Brian Langner, Antoine Raux, Alan W. Black, Maxine Eskenazi, and Alex Rudnicky. 2006. Online supervised learning of non-understanding recovery policies. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 170 –173, dec.

Dan Bohus. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*. Ph.D. thesis, Carnegie Mellon University.

Beth A. Hockey, Oliver Lemon, Ellen Campana, Laura Hiatt, Gregory Aist, James Hieronymus, Alexander Gruenstein, and John Dowding. 2003. Targeted Help for Spoken Dialogue Systems: intelligent feedback improves naive users' performance. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 147–154. Association for Computational Linguistics.

Stasinos Konstantopoulos, Athanasios Tegos, Dimitris Bilidas, Ion Androutsopoulos, Gerasimos Lampouras, Prodromos Malakasiotis, Colin Matheson, and Olivier Deroo. 2009. Adaptive natural language interaction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, EACL '09, pages 37–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theodora Koulouri and Stasha Lauria. 2009. A WOz framework for exploring miscommunication in HRI. In *Procs. of the AISB Symposium on New Frontiers in Human-Robot Interaction*, pages 1–8.

Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3&4):323–340, September.

Chris Mellish, Mick O'Donnell, Jon Oberlander, and Alistair Knott. 1998. An architecture for opportunistic text generation. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 28–37.

Verena Rieser, Ivana Kruijff-Korbayová, and Oliver Lemon. 2005. A Corpus Collection and Annotation Framework for Learning Multimodal Clarification Strategies. In *6th SIGdial Workshop on Discourse and Dialogue*.

David Schlangen and Raquel Fernández. 2006. Beyond Repair Testing the Limits of the Conversational Repair System. In *7th SIGdial Workshop on Discourse and Dialogue*.

Gabriel Skantze. 2003. Exploring human error handling strategies: Implications for spoken dialogue systems. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 71–76.

Dimitris Vogiatzis, Constantine D. Spyropoulos, Stasinos Konstantopoulos, Vangelis Karkaletsis, Zerrin Kasap, Colin Matheson, and Olivier Deroo. 2008. An affective robot guide to museums. In *Proceedings of the 4th International Workshop on Human-Computer Conversation, Bellagio, Italy*.

M Walker, C Kamm, and D Litman. 2000. Towards developing general models of usability with paradise. *Natural Language Engineering*, 6(3&4):363–377.

Teresa Zollo. 1999. A study of human dialogue strategies in the presence of speech recognition errors. In *Psychological Models of Communication in Collaborative Systems. Papers from the 1999 AAAI Fall Symposium (TR FS-99-03)*, pages 132–9. AAAI Press.

Processing Self-Repairs in an Incremental Type-Theoretic Dialogue System*

Julian Hough **Matthew Purver**
Interaction, Media and Communication group
School of Electronic Engineering and Computer Science
Queen Mary University of London
{julian.hough, matthew.purver}@eecs.qmul.ac.uk

Abstract

We present a novel incremental approach to modelling self-repair phenomena in dialogue, using the grammar and parsing mechanism of Dynamic Syntax (DS) to construct Type Theory with Records (TTR) record type representations incrementally in both parsing and generation. We demonstrate how a DS-TTR hybrid implementation when integrated into an incremental dialogue system can be exploited to account for the semantic processing of self-repair phenomena in a unified way and in line with psycholinguistic evidence.

1 Introduction

Self-repairs are too pervasive in human dialogue to be considered exceptional and they should be integral to any dialogue model, an insight from early Conversational Analysis work which revealed them to be systematic (Schegloff et al., 1977). The following are typical of the within-turn (first position), self-initiated type of self-repair often found in natural dialogue:

(1) Our situation is just [a little bit, + kind of the opposite] of that *(Switchboard)*

(2) [the interview was {...} + it was] alright *(Clark, 1996, p.266)*

(3) John goes to Paris [{uhh}+ from London] *(constructed example)*

*This work was partly supported by the RISER project EPSRC reference EP-J010383-1 and an EPSRC Doctoral Training Account scholarship for the first author. Thanks to the Semdial reviewers for their helpful comments.

For terminological and annotation purposes, following the disfluency-tagged Switchboard corpus, first position self-repairs will be discussed with reference to a division into a *reparandum* (the speech that is repaired, up to the repair point +), a possibly null *interregnum* (the filler words or pause between {}) and the following *repair* (the strings after the repair point + up to the closing square bracket). We also consider *extensions*- also called 'covert' (Levelt, 1989) or 'forward-looking' (Ginzburg et al., 2007) repairs- such as (3) which may not in fact function to alter the previous part of the utterance, but to extend it.

The formal model we describe here attempts to address two principal aspects of self-repair phenomena: firstly, in terms of cognitive processing, the semantic parsing and generation of self-repaired utterances is just as straightforward as for fluent utterances in dialogue (and in fact, in some domains semantic processing is aided (Brennan and Schober, 2001)); secondly, that the repaired material (reparandum) can be referred to in context, as in (2) above where the reparandum still needs to be accessed for the anaphoric use of 'it' to succeed, "leaving the incriminated material with a special status, but within the discourse context" (Ginzburg et al., 2007, p. 59).

2 Related work

Work on the processing of self-repair phenomena has not generally focused on the semantics and pragmatics of ongoing dialogue. Parsing approaches have tended to implement a *parse→string-edit→reparse* pipeline, which which takes disfluent inputs and returns ‘cleaned-up’ grammatical strings relative to a given grammar- this was done with a TAG transducer in the case of (Johnson and Charniak, 2004). In terms of psychological validity for dialogue the approach is questionable, as parts of an utterance cannot be removed from the hearer’s perceptual record, discounting the possibility of properly processing reparanda, as in example (2) above. McKelvie (1998) introduces a more explicit disfluency rule-based syntactic account, which instead of expunging ‘junk’ material, exploits *aborted* syntactic categories and provides optional rules for producing cleaned-up parses. However, again under the assumption that self-repair operates as a module outside the principal grammar, no method for obtaining the semantics of a self-repair is suggested.

Self-repair has received more attention from the generation (NLG) community, particularly as incremental NLG models were initially motivated by psycholinguistics, most notably Levelt (1989)’s influential modularization of speech production into distinct *conceptualization*, *formulation* and *articulation* phases. Following this, De Smedt showed how developing the syntactic component of the formulation phase in detail could give models of lexical selection and memory limitations (De Smedt, 1991) which could trigger syntactic revision and Neumann (1998) introduced reversible incremental parsing-generation processes to implement ambiguity detection and paraphrasing corrections. In conceptualization, Guhe (2007) modelled online modifications to pre-verbal messages that cause self-repair surface forms to be realized.

Albeit less psychologically motivated, Buß and Schlangen (2011) and Skantze and Hjalmarsson (2010) introduce self-repair generation strategies in incremental dialogue systems. Both systems make use of the Incremental Unit (IU) dialogue framework model (Schlangen and Skantze, 2009), which allows online revision of input and outputs between modules. Skantze and Hjalmarsson (2010) use string-based speech plans which may change dynamically during interaction with a user, allowing for changing ASR hypotheses, which could lead to the generation of a limited set of ‘covert’ (non-replacement extensions) and ‘overt’ self-repairs. The interactional benefits of the approach are clear, however the lack of incremental semantics and domain-general grammar makes scalability to more complex domains and integration with a parsing module difficult.

In terms of the dialogue semantics of self-repair, Ginzburg and colleagues (Ginzburg et al., 2007; Ginzburg, 2012) working within the KoS framework (Ginzburg, 2012) with Dialogue Gameboard (DGB) update mechanisms at its core, attempt to unify an account of self-repair and other-initiated repair by drawing the parallels between self-initiated editing phrases (interregna) and clarification requests (CRs) as cues for repair. They make an adjustment to KoS in allowing CRs and editing signals and their following corrections to occur mid-utterance, accommodating incrementality by allowing the DGB word-by-word updates to its PENDING component. They also suggest that Type Theory with Records (TTR) could be instrumental in enabling appropriate types for word-by-word semantic updates in their future work. However, while this provides a general dialogue model, the relationship of these updates to incremental parsing and generation processes is not made explicit.

3 Criteria for a unified account

The parsing, generation and dialogue semantics implementations of self-repairs have been slightly orthogonal, so a grammar which can provide a suitable semantic representation to capture the phenomena in both modalities within a dialogue context is lacking. We suggest that two requirements of a grammar to remedy this are *strong incremental interpretation* and *incremental representation* (Milward, 1991). Strong incremental interpretation is the ability to make available the maximal amount of information possible from an unfinished utterance as it is being processed, particularly semantic dependencies (e.g. a representation such as $\lambda x.like'(john', x)$ should be available after processing “John likes”). Incremental representation, on the other hand, is defined as a representation being available for each substring of an utterance, but not necessarily including the dependencies (e.g. having a representation such as $john'$ attributed to “John” and $\lambda y.\lambda x.like'(y, x)$ attributed to “likes” after processing “John likes”). These representations should become available immediately to connected modules, therefore requiring seamless integration with other dialogue semantics.[1]

Furthermore a record of *processing context* is required to be sufficiently detailed, and suitably structured, so that parsing, generation and dialogue management algorithms can access the material in the reparandum straightforwardly, as shown by example (2). This context must extend from the phonetic level to the conceptual level: Brennan and Schober (2001)'s experiments demonstrated self-repair can speed up semantic processing (or at least object reference) in a small visual domain of shape selection, where an incorrect object being partly vocalized and then repaired in the instructions (e.g. "the yell-purple square") yielded quicker response times from the onset of the target ("purple") than in the case of the equivalent fluent instructions (e.g. "the purple square"). This example will be addressed in section 6.

Given these requirements and the lacunae from previous work, in the remainder of the paper we present a type-theoretic incremental model of parsing, generation and context that addresses them, showing how a suitable grammar formalism and semantic representation (section 4) integrated into an incremental dialogue system (section 5) can account for parsing (section 6) and generating (section 7) self-repairs in a psycholinguistically plausible way.

4 Dynamic Syntax and TTR

Dynamic Syntax (DS) (Kempson et al., 2001) is an action-based and semantically oriented incremental grammar framework that defines grammaticality as parsability. The DS lexicon consists of *lexical actions* keyed to words, and also a set of globally applicable *computational actions*, both of which constitute packages of monotonic update operations on semantic trees, and take the form of IF-THEN action-like structures. For example, in DS notation, the lexical action corresponding to the word *john* has the preconditions and update operations in example (4): if the pointer object ($\diamondsuit$), which indicates the node being checked on the tree, is currently positioned at a node that satisfies the properties of the precondition then all the actions in the post-condition can be completed, these being simple monotonic tree operations.

(4)
$$
\begin{array}{ll}
\multicolumn{2}{l}{\textit{john:}} \\
\text{IF} & ?Ty(e) \\
\text{THEN} & \texttt{put}(Ty(e)) \\
 & \texttt{put}(\left[\ x_{=john}\ :\ e\ \right]) \\
\text{ELSE} & \texttt{abort}
\end{array}
$$

(5) $\overset{\textit{"John arrived"}}{\longmapsto}$

$$
\begin{array}{c}
\diamondsuit, Ty(t), \left[\begin{array}{ll} x_{=john} & : e \\ p_{=arrive(x)} & : t \end{array}\right] \\
\swarrow \qquad\qquad \searrow \\
\begin{array}{c} Ty(e), \\ \left[\ x =_{john}\ :\ e\ \right] \end{array}
\qquad
\begin{array}{c} Ty(e \rightarrow t), \\ \lambda r : \left[\ x1\ :\ e\ \right] \\ \left[\begin{array}{ll} x_{=r.x1} & : e \\ p_{=arrive(x)} & : t \end{array}\right] \end{array}
\end{array}
$$

In DS, the trees upon which actions operate represent terms in the typed lambda calculus, with mother-daughter node relations corresponding to semantic predicate-argument structure, with no independent layer of syntax represented. Tree nodes are typed, and can be either type-complete (e.g. $Ty(e)$) and decorated with a semantic formula, or have a requirement for a type (e.g. $?Ty(e)$). As can be seen in (5) above, recent DS variants (Purver et al., 2010) incorporate Type Theory with Records (TTR) (Cooper, 2005), with TTR *record types* decorating tree nodes, rather than simple atomic formulae.

Following Cooper (2005), each field in a record type is of the form $\left[\ l\ :\ T\ \right]$, containing a unique label l in the record type and a type T. Fields can be *manifest*, i.e. have a singleton type such as $\left[\ l\ :\ T_a\ \right]$ where T_a is the type of which only a is a member; here, we write this using the syntactic sugar $\left[\ l_{=a}\ :\ T\ \right]$. Fields can be *dependent* on fields preceding them (i.e. higher up in the graphical representation), e.g. the predicate type $\left[\ p_{=like(x,y)}\ :\ t\ \right]$, where x and y are labels in preceding fields. DS node semantic formulae are now taken to be record types, with the type of the final (i.e. lowest down) field corresponding to the $Ty()$ node type. Functions from record type to record type in the variant of TTR we use here employ paths, and are of the form $\lambda r : \left[\ l1\ :\ T1\ \right] \left[\ l2_{=r.l1}\ :\ T1\ \right]$, an example being the formula at the type $Ty(e \rightarrow t)$ node in tree (5) above, giving DS-TTR the required functional application capability: functor node functions are applied to their sister argument node's

[1] Recently, (Peldszus et al., 2012) show how incrementally integrating incremental syntactic and pragmatic processing can improve an interpreter module's performance.

formula, with the resulting β-reduced record type added to their mother node.[2]

In DS parsing, beginning with an axiom tree with a single node of requirement type $?Ty(t)$, parsing intersperses the testing and application of both lexical actions triggered by input words such as 4 and the execution of permissible (Kleene* iterated) sequences of computational actions, with their updates monotonically constructing the tree. Successful parses are sequences of action applications that lead to a tree which is complete (i.e. has no outstanding requirements on any node, and has type $Ty(t)$ at its root node as in (5)). The DS notion of incrementality is two-fold, in that action sequences monotonically extend the trees, and that these sequences are maximally applied on a word-by-word basis.

Here we modify the traditional DS parsing and generation model by allowing the compilation of TTR formulae for *partial* trees in addition to complete ones. This is achieved through a simple tree-compiling algorithm which decorates terminal nodes with record types containing underspecified variables of the appropriate type, then applies functional application between sister nodes to compile a β-reduced record type at their mother node, continuing in bottom-up fashion until a record type is compiled at the root (see (Hough, 2011) for details). The modification means the DS-TTR model now meets the criteria of strong incremental interpretation, as maximal record types represent all possible dependencies made available as each word is processed.

4.1 DS-TTR generation as parsing

As Purver and Kempson (2004) demonstrate, an incremental DS model of surface realization can be neatly defined in terms of the DS parsing process and a *subsumption check* against a *goal tree*. The goal tree input is a complete and fully specified DS tree such as (5), and the generation of each word consists of attempting to parse each word in the lexicon to extend the trees under construction in the parse state. Partial trees are checked for suitability via goal tree subsumption, with unsuitable trees and their parse paths removed from the generator state. The DS generation process is word-by-word incremental with maximal tree representations continually available, and it effectively combines lexical selection and linearization into a single action due to the word-by-word iteration through the lexicon. Also, self-monitoring is inherently part of the generation process, as each word generated is parsed. However, this model requires fully structured *trees* as input, problematic for a dialogue manager.

Here, though, with incremental representations now available through the tree compiling mechanism as described above, a modification can be made by replacing the goal tree with a *TTR goal concept*, which can take the form of a record type such as:

(6) $$\begin{bmatrix} x1_{=Paris} & : & e \\ x_{=john} & : & e \\ p1_{=to(x1)} & : & t \\ p_{=go(x)} & : & t \end{bmatrix}$$

Consequently, the tree subsumption check in the original DS generation model can now be characterized as a TTR subtype relation check between the goal tree and the compiled TTR formulae of the trees in the parse state. A definition for the check, adapted from Fernández (2006, p.96), is defined in (7).

(7) *Subtype relation check*:
For record types $p1$ and $p2$, $p1 \sqsubseteq p2$ holds just in case for each field $\begin{bmatrix} l : T2 \end{bmatrix}$ in $p2$ there is a field $\begin{bmatrix} l : T1 \end{bmatrix}$ in $p1$ such that $T1 \sqsubseteq T2$, i.e. iff any object of type $T1$ is also of type $T2$. This relation is reflexive and transitive.

The advantage of this move is that for the logical input to generation a goal tree no longer needs to be constructed from the grammar's actions, so the dialogue management module need not have full knowledge of the DS parsing mechanism and lexicon. An example successful generation path can be seen in Figure 1,[3] showing how the maximal TTR record type for each tree is continually available.

[2] For functional application and record type extension (concatenation), which is required in DS grammar for merging the formulae at the top of LINKed tree structures, *relabelling* is carried out when necessary in the record types in the way described by Cooper (2005) and Fernández (2006).

[3] The incremental generation of "john arrives" succeeds as the successful lexical action applications at transitions $\boxed{1} \mapsto \boxed{2}$ and $\boxed{3} \mapsto \boxed{4}$ are interspersed with applicable computational action sequences at transitions $\boxed{0} \mapsto \boxed{1}$ and $\boxed{2} \mapsto \boxed{3}$, at each stage passing the subtype relation check with the goal (i.e. the goal is a subtype of the top node's compiled record type), until arriving at a tree that *type matches* in $\boxed{4}$.

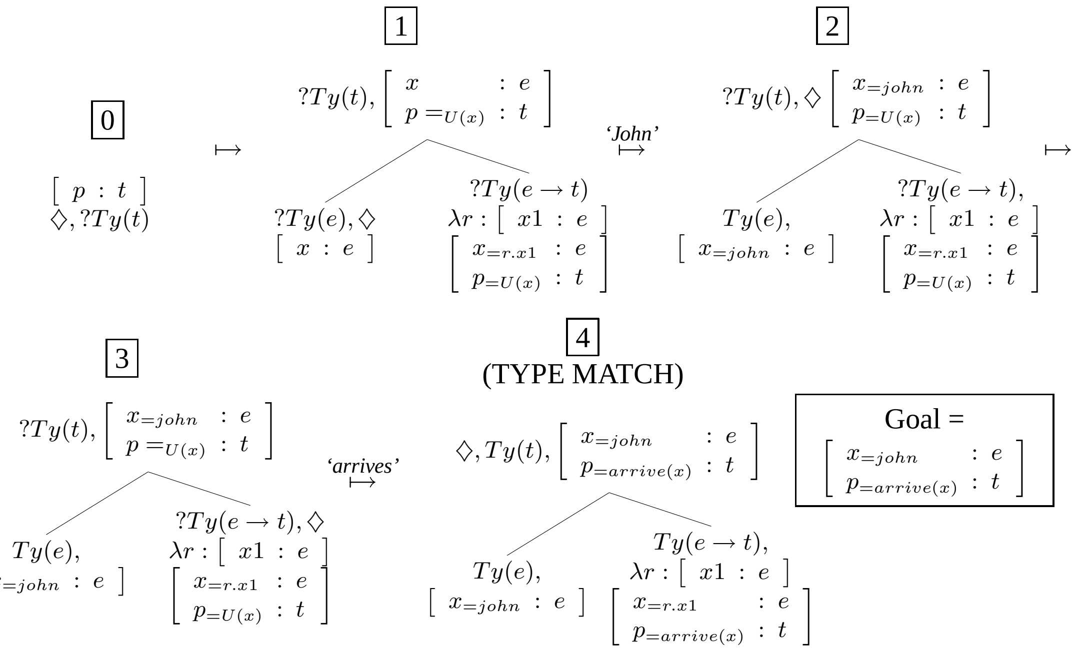


Figure 1: Successful generation path in DS-TTR

Another efficiency advantage is that subtype checking can also reduce the computational complexity of lexicalisation through pre-verbal lexical action selection, removing the need to iterate through the entire lexicon on a word-by-word basis. A sublexicon $SubLex$ can be created when a goal concept $GoalTTR$ is inputted to the generator by searching the lexicon to select lexical actions whose TTR record type is a valid supertype of $GoalTTR$.

5 Incremental DS-TTR parsing and generation in DyLan

In order to meet the criteria of a continuously updating contextual record, we implement DS-TTR parsing and generation mechanisms in the prototype `DyLan` dialogue system[4] within Jindigo (Skantze and Hjalmarsson, 2010), a Java-based implementation of the incremental unit (IU) dialogue system framework (Schlangen and Skantze, 2009). As per Schlangen and Skantze (2009)'s model, there are input and output IUs to each module, which can be `add`ed as edges between vertices in module buffer graphs and become `commit`ted should the appropriate conditions be fulfilled, a notion which becomes important in light of hypothesis change and repair situations. Dependency relations between different graphs within and between modules can be specified by *groundedIn* links (see (Schlangen and Skantze, 2009) for details).

The `DyLan` interpreter module (Purver et al., 2011) uses Sato (2011)'s insight that the context of DS parsing can be characterized in terms of a Directed Acyclic Graph (DAG) with trees for nodes and DS actions for edges. The module's state is characterized by three linked graphs:

- *input:* a time-linear word graph posted by the ASR module, consisting of word hypothesis edge IUs between vertices W_n
- *processing:* the internal DS parsing DAG, which adds parse state edge IUs between vertices S_n *groundedIn* the corresponding word hypothesis edge IU
- *output:* a concept graph consisting of domain concept IUs (TTR record types) constructed between vertices C_n, *groundedIn* the corresponding path of edges in the DS parsing DAG

In the generation module, the architecture is the inverse of interpretation given the input of TTR goal concepts:

- *input:* the concept graph has goal concept IU edges (TTR record types) between vertices GC_n posted by the dialogue manager

[4] Available from http://dylan.sourceforge.net/

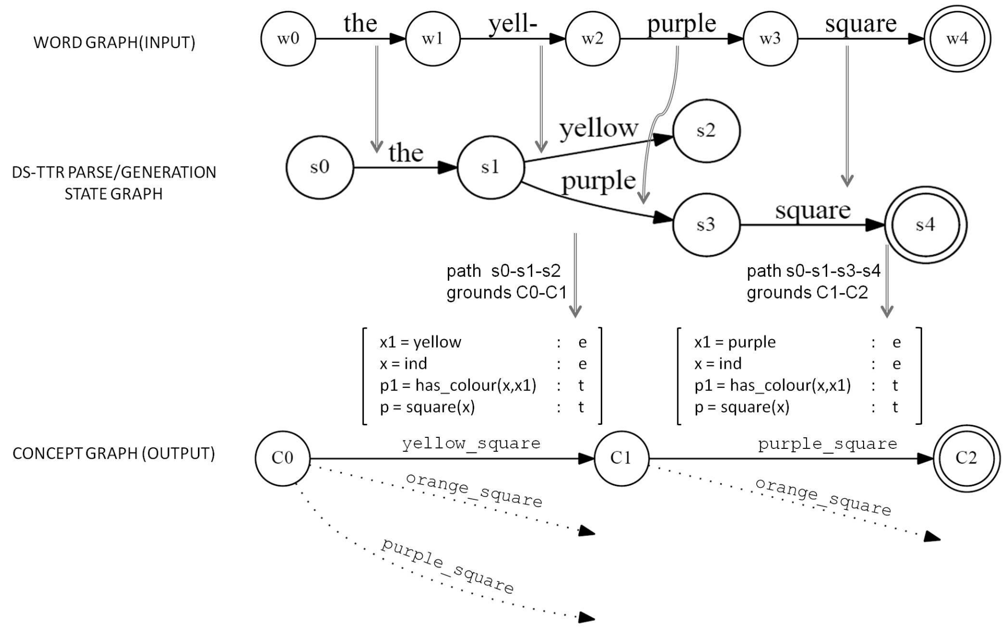


Figure 2: Incremental DS-TTR parsing of a self-repair. Revoked edges indicating failed search paths are dotted. Inter-graph *groundedIn* links go from top to bottom.

- *processing:* the DS parsing graph (shared with the interpreter module's graph) is incrementally constructed word-by-word by parsing the lexical actions in the sublexicon and subtype checking the result against the current goal concept (see section 4.1)
- *output:* the word graph's edges are `add`ed to the output buffer during word-by-word generation, and `commit`ted when they are *groundedIn* DS parsing graph paths that form part of a valid generation path *type matched* with the goal concept (as in Figure 1).

6 Parsing self-repairs

Interpretation in `DyLan` follows evidence that dialogue agents parse self-repairs efficiently and that repaired material is given special status but not removed from the discourse context. To simulate Brennan and Schober (2001)'s experimental findings described in section 3, we demonstrate a self-repair parse in Figure 2 using a domain of three domain concepts, `yellow_square`, `purple_square` and `orange_square`, each with a distinct record type. When "yell-" is processed, the word hypothesizer adds the edge 'yellow', which in turn is parsed, returning a TTR record type. Search is initiated for domain concepts in a subtype relation to it, in this case finding a valid subtype in the concept `yellow_square`- when matched it is moved from the domain concepts to the concept graph's active edge. The following failure to interpret 'purple' forces a repair under the definition in 8 below:

(8) **Repair** IF from parsing word W there is no edge SE_n able to be constructed from vertex S_n (no parse) or if no domain concept hypothesis can be made through subtype relation checking, `repair`: parse word W from vertex S_{n-1}. Should that parse be successful `add` a new edge to the top path, without removing any `commit`ted edges beginning at S_{n-1}.

This does not remove the initially matched concept IU at edge $C0$-$C1$, but forces another matching process to add a successor edge. The consequent subtype-checking operation is then limited to just the concepts `purple_square` and `orange_square`, finding a type match in the former. While this trivially reduces the subtype checking iteration process here for illustrative purposes, with a bigger domain this could remove many concepts (i.e. all of those that are subtypes of the incriminated parse path's current record type).

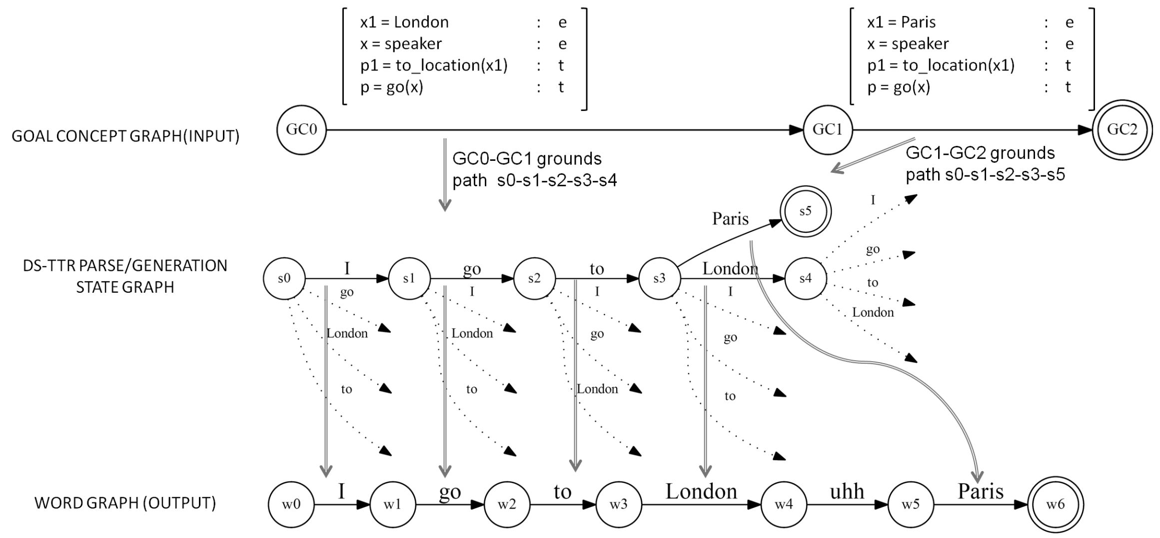


Figure 3: Incremental DS-TTR generation of a self-repair upon goal concept change. Type-matched record types are double-circled nodes. Inter-graph *groundedIn* links go from top to bottom.

This strategy will also allow the parsing of (2) "the interview was.. it was alright", with the correct reference resolution of 'it': any committed preceding edge on the word hypothesis graph can be accessed (i.e. any word/partial word heard in the user's speech stream), as can its corresponding *groundedIn* DS-TTR parse graph edge IU, so the TTR formula for 'the interview' is accessible and DS anaphora mechanisms using context may run as normal.

While the rule in (8) will only allow the parsing of replacement type self-repairs, in our prototype dialogue system this can be triggered not only by syntactic disfluency but also by pragmatic infelicity. For example, if the user were to say "I pick the yellow square or rather the blue square", which may be parsable in the DS grammar without backtracking, the mechanism will still work in the same way because in our micro-domain there is no available concept that represents the user selecting both the yellow and blue squares simultaneously in one turn. Work is also under way to lexicalise editing signals in terms of their effect on DS parsing context.

7 Generating self-repairs

In `DyLan`'s generation module, whose processing is driven by parsing as described in section 4.1, the parsing `repair` function defined in (8) will operate if there is no resulting word edge output after a generation cycle to produce the next word. This will be triggered by a change in goal concept during generation. As per parsing, in repair the generation algorithm continues backtracking by one vertex at a time in an attempt to extend the DS DAG until successful, as can be seen in Figure 3 with the successful backtrack and parse of 'Paris' resulting in successful subsumption to the new goal concept. The time-linear word graph continues to extend but with the repair's edges *groundedIn* different paths of the parse DAG to those which ground the reparandum.[5] Our protocol is consistent with Shriberg and Stolcke (1998)'s empirical observation that the probability of retracing N words back in an utterance is more likely than retracing from N+1 words back, making the repair as local as possible.

Another type of self-repair, *extension*, such as example (3) above, is dealt with straightforwardly in our generation module. For these covert repairs, the incoming goal concept must be a subtype of the one it replaces, and so the DS parser can induce monotonic growth of the matrix tree through LINK adjunction (Kempson et al., 2001), resulting in subtype extension of the root TTR record type. Thus, a change in goal concept during generation will not always put demands on the system to backtrack, such as in generating the fragment after the pause in "John goes to Paris. . . from London". It is only

[5]The previously committed word graph edge for 'London' is not revoked nor is its *groundedIn* parse graph edge, following our parsing algorithm and the principle that has been in the public record and hence should still be accessible.

at a semantics-syntax mismatch, where the revised goal TTR record type does not correspond to a permissible extension of a DS tree in the parsing DAG where overt repair will occur. In contrast to Skantze and Hjalmarsson (2010)'s string-based *speech plan* comparison approach, there is no need to regenerate a fully-formed string from a revised goal concept and compare it with the string generated thus far. Far from a phonetic form deletion account, self-repair in `DyLan` is driven by attempting to extend existing parse paths to construct the new target record type, *retaining* all the semantic representation and the procedural context of actions already built up in the generation process to avoid the computational demand of constructing semantic representations from afresh. This way a unified mechanism for modification and extension repairs is possible.

8 Conclusion and Future Work

We have presented a framework for parsing and generating word-by-word incrementally using a hybrid grammar of Dynamic Syntax and Type Theory with Records (DS-TTR) which has been implemented in the `DyLan` dialogue system, utilising the mechanisms of the Incremental Unit framework. `DyLan` provides a preliminary model of the parsing and generation of self-repair in line with psycholinguistic evidence of preference for locality and the availability of access to the semantics of repaired material.

In terms of development, while our model currently covers replacement type repairs and extensions, there is potential for expansion to insertion type repairs such as "Peter went swimming with Susan... or rather surfing, yesterday".[6] The use of a DS-TTR parsing context DAG constructed by the utterance so far could again be used to resolve these repairs, in this case by reusing preceding action edges in the spirit of the recent DS account of verb phrase ellipsis (VPE) (Kempson et al., forthcoming). Schematically, the repair mechanism would work in a similar way to the resolution of the VPE in "[Peter went] swimming [with Susan] and $_{<Ai-Aj>}$ surfing $_{<Ak-Al>}$ yesterday", where $_{<Ai-Aj>}$ is the sequence of action edges used in the construction of the DAG triggered by the words "Peter went" and $_{<Ak-Al>}$ is the sequence for "with Susan", both of which are re-used either side of the actions triggered by "surfing". The regeneration (Kempson et al., forthcoming) rule operating over DS context makes this possible here, enabling the parser (and generator) to take a sequence of actions from context and re-use them, provided that they were triggered by the same type-requirement as is imposed on the node currently under development in the tree being constructed in the DAG. The difference between the VPE and repair mechanisms would lie in the commitment status of the TTR formulae constructed by these action paths: in VPE the record types yielded from both construction paths would be committed to the dialogue manager, whilst in repair, although the reparandum's path remains accessible to the parsing and generation modules, only the repair's resultant TTR record type would be committed. This account is yet to be fully fledged out for these repairs, particularly the consideration of how editing terms would be interpreted appropriately, and the computational demands for the on-line resolution of the repair, however the potential for extending the self-repair model provided by the DS characterization of context is clear.

Other future work planned includes investigating the lexical semantic structure that TTR record types may offer for modelling type dependencies between reparanda and repairs.

[6] Many thanks to the anonymous reviewer who provided the example and highlighted the importance of this.

References

S.E. Brennan and M.F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech* 1. *Journal of Memory and Language*, 44(2):274–296.

O. Buß and D. Schlangen. 2011. Dium–an incremental dialogue manager that can produce self-corrections. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, pages 47–54, Los Angeles, California, September.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

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

Koenraad De Smedt. 1991. Revisions during generation using non-destructive unification. In *Proceedings of the Third European Workshop on Natural Language Generation*, pages 63–70.

Raquel Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College London, University of London.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2007. Unifying self- and other-repair. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue (DECALOG)*.

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

Markus Guhe. 2007. *Incremental Conceptualization for Language Production*. NJ: Lawrence Erlbaum Associates.

Julian Hough. 2011. Incremental semantics driven natural language generation with self-repairing capability. In *RANLP 2011 Student Research Workshop*, pages 79–84, Hissar, Bulgaria.

Mark Johnson and Eugene Charniak. 2004. A tag-based noisy channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 Semantic Theory*. 2nd edition.

W.J.M. Levelt. 1989. *Speaking: From intention to articulation*. MIT Press.

David McKelvie. 1998. The syntax of disfluency in spontaneous spoken language. In *HCRC Research Paper, HCRC/RP-95*.

David Milward. 1991. *Axiomatic Grammar, Non-Constituent Coordination and Incremental Interpretation*. Ph.D. thesis, University of Cambridge.

Günter Neumann. 1998. Interleaving natural language parsing and generation through uniform processing. *Artificial Intelligence*, 99:121–163.

Andreas Peldszus, Okko Buß, Timo Baumann, and David Schlangen. 2012. Joint satisfaction of syntactic and pragmatic constraints improves incremental spoken language understanding. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–523, Avignon, France, April. Association for Computational Linguistics.

Matthew Purver and Ruth Kempson. 2004. Incremental context-based generation for dialogue. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG04)*, number 3123 in Lecture Notes in Artifical Intelligence, pages 151–160, Brockenhurst, UK, July. Springer.

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 *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, 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 *Proceedings of the 9th International Conference on Computational Semantics*, pages 365–369, Oxford, UK, January.

Yo Sato. 2011. Local ambiguity, search strategies and parsing in Dynamic Syntax. In E. Gregoromichelaki, R. Kempson, and C. Howes, editors, *The Dynamics of Lexical Interfaces*, pages 205–233. CSLI.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece, March. Association for Computational Linguistics.

Elizabeth Shriberg and Andreas Stolcke. 1998. How far do speakers back up in repairs? A quantitative model. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2183–2186.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 1–8, Tokyo, Japan, September. Association for Computational Linguistics.

Modelling Strategic Conversation: the STAC project

N. Asher [1], A. Lascarides [2], O. Lemon [3], M. Guhe [2], V. Rieser [3]
P. Muller [4], S. Afantenos [4], F. Benamara [4], L. Vieu [1]
P. Denis [5], S. Paul [4], S. Keizer [3], and C. Dégremont [4]
[1] CNRS, Laboratoire IRIT
[2] University of Edinburgh
[3] Heriot-Watt University
[4] Université Paul Sabatier, Toulouse
[5] Equipe MOSTRARE, INRIA LNE – LIFL

Abstract

The STAC project will develop new, formal and robust models of non-cooperative conversation, drawing from ideas in linguistics, philosophy, computer science, and economics. The project brings a state of the art, linguistic theory of discourse interpretation together with new data-driven models of agent interaction and strategic decision making. Here we discuss the project's linguistic underpinnings, and the conceptual and empirical challenges the project faces. We also describe the project's current data collection activities.

1 Introduction

An assumption of all implemented dialogue systems and almost all formal work on discourse is that agents are fully cooperative. What this means is that agents adhere to principles such as: normally one believes what one says (e.g. (Grosz and Sidner, 1990)) and that one normally tries to help one's interlocutors achieve their goals. The latter in turn requires speakers to adopt shared intentions; in other words, their preferences are fully aligned, and they are *sincere* —they believe what they say. Consequently, all dialogue systems to date are limited to domains where such assumptions are sustainable, such as tourist information. But there are many scenarios where this level of cooperativity doesn't apply: for example, dialogues involving complex negotiations (Traum, 2008), or political debate (Lipman and Sippi, 1995). In a dialogue from the Settlers of Catan game below, (1b) is true but misleading because it implicates that B doesn't have rock.

(1) a. A: Do you have rock?
b. B: I've got lots of wheat [in fact, B has a rock]
c. A: I'll give you 2 clay for a rock
d. B: How about 2 clay for a wheat?
e. A: I'll give 1 clay for 3 wheat
f. B: OK, it's a deal.

Nevertheless, here cooperativity has not broken down entirely: (1b) supplies an (indirect) answer to (1a), and so in contrast to an assertion such as *I won't answer* it meets at least one goal that is associated with asking a question. Similarly, A's assertion (1e) attends to B's underlying goal in uttering (1d), of obtaining clay. We also note here that such dialogue contributions describe the preferences of individuals, so that modelling complex preferences will be a key feature of dialogue state representation for STAC.

2 Data collection: Settlers of Catan

We are currently collecting non-cooperative dialogue data using an online version of the popular board game "Settlers of Catan" (see figure 1). Negotiation dialogues are a critical part of the game, and information hiding and deception are observed in the data. The original JSettlers interface was developed by (Thomas and Hammond, 2002), and we have modified it to include a chat tool whereby players' trading dialogues are being collected (Guhe and Lascarides, 2012). An annotation scheme for non-cooperative negotiation dialogues is being developed in the project.

3 Project components

3.1 Modelling Preferences

Just as the dynamic semantics of SDRT treats an utterance as a relation between information states, we will treat utterances as relations or transitions between preference states. These states reflect the

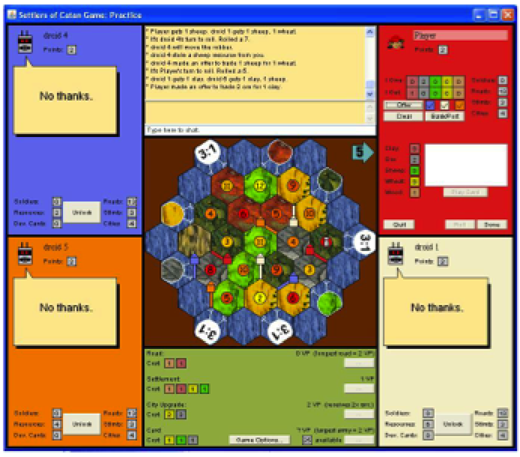


Figure 1: The adapted J-Settlers interface

structure and logical dependencies among the various factors that influence agent behaviour. *Conditional preference* (CP) nets (Boutilier et al., 2004) provide a computationally effective and highly compact representation for expressing and reasoning with preferences over large sets of features, and we will use this in our model. CP-nets provide an effective way to handle the fact that dialogue often reveals complex preferences, incorporating dependencies between features. STAC will determine algorithms for uncovering preferences from conversation.

3.2 Modeling Non-Cooperative Dialogue

Segmented Discourse Representation Theory (SDRT) has a well articulated theory of dialogue, which provides a clear and formal interaction between attributions of attitudinal states and discourse contributions (Asher and Lascarides, 2003). The dynamics of SDRT allows us to constrain agent modelling, restricting search over actions and preferences (Asher and Lascarides, 2008). However, SDRT's cognitive logic, as detailed in (Asher and Lascarides, 2003), is a static, BDI logic that fails to reflect the structural complexity of decision problems, and it has nothing to say about less than completely cooperative and infallible agents. Work in STAC is replacing this cognitive model to address these shortcomings.

3.3 Statistical models and Machine Learning

We are also exploring how reinforcement learning (RL) — a statistical planning method for acquiring optimal dialogue policies (see e.g. (Rieser and Lemon, 2011)) — can be used to learn optimal *strategic* dialogue policies. A new challenge for RL is to work in non-cooperative domains such as resource negotiation in Settlers, where an agent may not be fully honest when expressing their preferences. This type of partial observability falls outside the scope of current Partially Observable Markov Decision Process (POMDP) approaches to dialogue, which focus on uncertainty derived from speech recognition errors. Finding a suitable generalization of the POMDP framework to handle such data is an important challenge for the STAC project.

4 Future Work

As well as using the Settlers domain, the STAC project is also exploring data from debating dialogues (Lipman and Sippi, 1995). Please see `http://www.irit.fr/STAC/`

Acknowledgments

The research leading to this work has received funding from the European Community (FP7/2011-16) under grant agreement no. 269427 (STAC).

References

Asher, N. and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Asher, N. and A. Lascarides. 2008. Commitments, beliefs and intentions in dialogue. In *Proc. of the Workshop on the Semantics and Pragmatics of Dialogue*, pages 35–42.

Boutilier, C., R. Brafman, C. Domshlak, H. Hoos, and D. Poole. 2004. CP-nets : A Tool for Representing and Reasoning with Conditional *Ceteris Paribus* Preference Statements. *Journal of Artificial Intelligence Research*, 21:135–191.

Grosz, B. and C. Sidner. 1990. Plans for discourse. In *Intentions in Communication*. MIT Press, pages 365–388.

Guhe, Markus and Alex Lascarides. 2012. Trading in a multiplayer board game: Towards an analysis of non-cooperative dialogue. In *Proc. 34th Conference of the Cognitive Science Society*.

Lipman, B. and D. Sippi. 1995. Robust inference in communication games with partial provability. *Journal of Economic Theory*, 66:370–405.

Rieser, Verena and Oliver Lemon. 2011. *Reinforcement Learning for Adaptive Dialogue Systems*. Theory and Applications of Natural Language Processing. Springer.

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

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

Toward a Mandarin-French Corpus of Interactional Data

Helen K.Y. Chen[1,2], Laurent Prévot[1], Roxane Bertrand[1], Béatrice Priego-Valverde[1], Philippe Blache[1]

(1) LPL, CNRS & Aix-Marseille Université
5 Ave Pasteur
13604 Aix en Provence-France

(2) Dept. of Chinese and Bilingual Studies
The HK Polytechnic University
Hung Hom, KLM, Hong Kong

FirstName.LastName@lpl-aix.fr

1. Introduction

This paper provides a preliminary description of the construction of an audio-video speech corpus of face-to-face Mandarin interaction. The corpus consists of 5 hours of two-party, face-to-face, spontaneous Mandarin interaction. Following the construction of the *Corpus of Interactional Data* (CID), an audio-video corpus in French recorded and processed at the Laboratoire Parole & Langage (LPL), Aix-Marseille Universités (Bertrand et al., 2008), the current project proposes the construction of a Mandarin speech corpus that will be annotated, based on a *multimodal* perspective, at various linguistic levels including prosody, morphology, syntax, as well as discourse and non-verbal representations such as gestures. The objective of building such a corpus is to provide a speech resource annotated with wealthy and detailed information at each linguistic level. The eventual goal is to facilitate analyses of conversational interaction from a multimodality perspective.

2. Design of the corpus – the protocol

It is preliminarily proposed that the corpus will consist of 5 segments of two-party, face-to-face Mandarin conversational recordings. Each recording lasts about 1 hour. Thus at the initial stage of the creation of the corpus, the total recording time will add up to 5 hours. The eventual goal is to create a corpus of Mandarin spontaneous conversations of at least 10 hours.

2.1 The subjects

5 males and 5 females are involved in the recordings of face-to-face interaction. All speakers are native speakers of Taiwanese Mandarin. Some of the speakers have grown up bilingually speaking also Taiwanese. Also some speakers may have worked in the same lab where the recording took place and are familiar with each other and the recording environment.

2.2 The task

The participants are invited to join the experiment, in which they are instructed to "have a chat" with another speaker. There are 3 sets of guidelines provided randomly to the participants prior to each recording session: 1. the participants may be instructed to talk about the most difficult things they've encountered while living in France; 2. the participants may be instructed to talk about one of the unusual things s/he has encountered during a recent trip; 3. the participants may not be given any instruction about what to talk about and simply started the conversation on their own. The reason to provide a guideline (especially for guideline 1 and 2) is to enable speakers to be engaged in the interaction rather quickly as soon as the recording starts. Although the speakers may be provided with one of the 3 guidelines as the initial topic for interaction, there was no further instruction about how long the topic should last. The speakers are free to switch to other topics as the interaction carries on.

2.3 The setting of the recordings

Following the original French CID setting (Bertrand et al., 2008), all recording session took place in a soundproof chamber at LPL. The two speakers in each session sat side-by-side and slightly tilted towards each other. During the recording, each subject wore a headset and the voice from each speaker was recorded onto a separated track. As result, the optimal quality of sound files of the spoken data can be obtained for the purpose of detailed annotations on the phonetic and prosodic levels. Moreover, the recordings with two separated sound tracks have the advantage of allowing a more detailed analysis on the content in the overlapped sequences (Bertrand et al., 2008). In addition to the specific setting for the sound recording, subjects were also filmed in long and fixed shot. The video recordings otherwise provide data for non-verbal cues such as gestures.

2.4 The characteristics of the corpus

As result of the aforementioned experimental design and settings, the conversations recorded for the CID corpus is presented as closely simulating the data of naturally occurred, face-to-face interaction. The dialogues of the current corpus resemble daily Mandarin interactions and can serve as a speech resource with rich information on turn-taking and sequential organizations of conversation (Sacks et al., 1974). Such characteristic of the corpus can contribute to the further analysis of the interactions between Mandarin speakers from an interesting range of perspectives: the correlation between conversational interaction and sound realization (such as at the phonetic or prosodic levels, see Chen 2011), the examination of interaction between the speakers in terms of syntactic constructions or semantic/pragmatic implicatures, and finally, but not the least, the non-verbal cues such as gestures.

3. Levels of transcriptions and annotations

The data will be transcribed and then annotated at various linguistic levels. The following describes the process of transcription and various levels of annotations proposed.

Following Blache et al. (2009), prior to the transcriptions the data will be pre-processed by an automatic segmentation into blocks of sound stream by silent pauses of 200ms. The purpose of the segmentation is to facilitate further orthographic and phonetic/phonological transcription, as well as the alignment of the signals and annotations. The segmented data will then be transcribed orthographically (by using standard Hanyu Pinyin). Additional phonetic and/or phonological transcriptions may be added later on.

At the prosodic level, the annotation scheme distinguishes two levels: a higher level of intonational phrases; and a lower one of interactionally related prosodic cues: including duration, cut-off, lengthening, special voice quality (such as laryngealized voice). For the lower level of prosodic annotations the unit will be the tokens.

Concerning discourse and interaction, we will start from the Turn Constructional Unit (TCU). Following Schegloff (2007), a TCU corresponds to an action in interaction, i.e. a question, an answer, a request, etc. It should be noted that a TCU does not necessarily correspond to a complete sentence; it can correspond to either a lexical item (as a reactive token), or a TCU may consist of more than one sentence. The discourse/interaction information will be annotated by TCU, with notations about the specific action that consists of the TCU: i.e., a question/answer pair, a request or refusal to the request, etc. Moreover a parallel annotation of disfluencies will identify the *reparandum, reparans* and *break interval* in the spirit of Shriberg (1994) (see also Pallaud, 2006, Chen, 2011).

Finally, with regard to the non-verbal gestures, 5 types of gestures will be identified, following McNeill (1992), as well as Chui (2003): *iconic, metaphoric, deictic, spatial* gestures and *beats*. Each type of gesture will be further marked by its *preparation*, *stroke*, and *retraction* (McNeill 1992).

4. Contribution of the corpus

In addition to obvious benefits of such a multimodal richly annotated corpus, the perfect replication of the CID experimental setting will allow for systematic cross-linguistic studies. In addition, the Mandarin CID corpus will be closely related to the Mandarin Conversational Dialogue Corpus (MCDC) (Tseng, 2004). However the CID corpora will be perfectly comparable while the MCDC is different in terms of the experimental settings (e.g the speakers did not know each other when doing the recoding.) Comparisons between these corpora would be nevertheless worthy further exploration.

References

Bertrand et al. (2008). "Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle." *Traitement automatique des langues* (TAL), vol. 49, no. 3. 105-134.

Blache et al. (2009). "Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project." *Multimodal Corpora*: *From Models of Natural Interaction to Systems and Applications*, Springer.

Chen, K. (2011). *Sound Patterns in Mandarin Recycling Repair*. Ph.D. dissertation. University of Colorado at Boulder.

Chui, K. (2003). Categorization of Gestures in Communication. *Form and Function*: *Linguistic Studies in Honor of Shuanfan Huang*. 105-129. Taipei: Crane.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. Berkeley, CA: University of California at Berkeley.

Sacks et al. (1974). "A simplest systematics for the organization of turn-taking for conversation." *Language* 50. 696–735.

Pallaud, B. (2006). Une base de données sur les troncations involontaires de mots en français parlé. *Travaux Interdisciplinaires de Parole et Langage* (TIPA), no. 25. 173-184.

Schegloff, E. (2007). *Sequence organization in interaction: A primer in conversation analysis I.* Cambridge & New York: Cambridge University Press.

Tseng, S-C. (2004). Processing Spoken Mandarin Corpora. *Traitement automatique des langues*. Special issue: Spoken corpus processing 45.89-108.

A model of intentional communication: AIRBUS (Asymmetric Intention Recognition with Bayesian Updating of Signals)

J. P. de Ruiter and Chris Cummins
Bielefeld University

Abstract

The rapid and fluent nature of human communicative interactions strongly suggests the existence of an online mechanism for intention recognition. We motivate and outline a mathematical model that addresses these requirements. Our model provides a way of integrating knowledge about the relationship between linguistic expressions and communicative intentions, through a rapid process of Bayesian update. It enables us to frame predictions about the processes of intention recognition, utterance planning and other-repair mechanisms, and contributes towards a broader theory of communication.

Introduction

The ability to communicate effectively and flexibly with other humans is one of our species' most impressive cognitive capacities. However, there are very few comprehensive theories that aim to address this capacity, and those that do are often sketchy and fail to capture the essential and unique aspects of human communication.

Most notably among these, Shannon's (1948) mathematical theory of signal transmission is of limited use in modeling human-human communication. This model assumes that the encoder function that the sender uses to convert a message into a signal is the inverse of the decoder function that the receiver uses to reconstruct the message from the signal. This is not descriptively adequate for human communication, whose complex many-to-many mappings sometimes break down, resulting in miscommunication. The influential recent Interactive Alignment model (Pickering and Garrod 2004) implicitly assumes even similar encoding and decoding functions, namely the identity function.

From a more linguistic perspective, Grice's (1957) theory of meaning provides a very concise definition of what constitutes (intentional) communication, but is atheoretic as to how this process is accomplished. Research in this tradition encounters the daunting complexities, and potential infinite regress, associated with the recognition of mutual knowledge or *common ground* (Stalnaker, 1978; Clark and Marshall, 1981). Reductionist approaches to this problem are motivated by the intuition that full common ground processing is implausible given the speed and efficiency of typical dialogue. The immediacy of turn-taking (Stivers et al. 2009) and back-channel responses (Yngve 1970) speak to the need for rapid online heuristics that enable hearers to identify the general nature of the speaker's communicative intention or illocution.

The absence of models of human communication that address these competing concerns is keenly felt. Here we propose a mathematical model of communication that crucially relies upon the use of *shared conventions* to achieve efficiency, and that applies a form of Bayesian updating to address the many-to-many mapping problem. Rather than attempting to apply machine learning techniques such as POMDP to learn optimized mappings from utterances to appropriate responses in one fell swoop, we focus on the more tractable problem of recognizing the category of utterance involved. This enables us to consider the full range of different communicative contexts without succumbing to unsolvable complexity in the case of infinitely productive human language.

In the following we outline the technicalities of the model and discuss some of its implications.

Outline specification of the model

The AIRBUS model takes a signal as its input and calculates the corresponding intention. The model assumes a finite, predefined set of communicative intentions. It has access to three forms of information: a convention database C, which specifies the probability of communicative intentions given a certain signal; a likelihood database L, specifying the probability of signals given a certain communicative intention, and a set of prior probabilities E as to the communicative intention, conditioned by the social and discourse context.

The operation of the model consists of updating the prior probabilities E in the light of a new incoming signal, taking into account the information in C and L. We propose the following stages of update. Given a new signal s, the model examines whether there is an entry in C corresponding to the signal s. If so, the probabilities in this entry are averaged with the probabilities in E, creating a set of revised probabilities R. R is then treated as a prior and subjected to Bayesian update in the light of L. The resulting probability distribution over I is used to infer the speaker's intention. This process cycles as the signal continues and further convention-bearing units are transmitted.

Within this model, we can measure the success or the usefulness of a communicative act by considering the extent to which it reduces the hearer's uncertainty as to the speaker's intention. Following Shannon (1948), we can measure this by considering the entropy of the prior and posterior probability distributions over the possible intentions in I. We propose that the hearer commences the planning of a response when the entropy is low enough.

Separately, we propose that repair mechanisms are activated if there is too large a difference between prior and posterior distributions: that is, if the hearer's understanding of the speaker's intention is radically altered during the update process. A large difference would suggest disalignment between speaker and hearer, and the possible need for explicit repair negotiation. We can measure this difference using Kullback-Leibler divergence, a standard measure of relative entropy, and posit that sufficiently high K-L divergence triggers explicit repair.

Discussion

The model outlined above provides a rapid means to infer communicative intentions. It posits a powerful decoding process, using the hearer's knowledge about both directions of the relationship between signal and intention to draw pragmatic conclusions about the speaker's intended meaning. Moreover, by its use of probability distributions rather than categorical rules, the model is able to handle improbable events gracefully. Relative entropy allows us to predict when the model does break down to such an extent that explicit negotiation is required.

In this brief sketch we have necessarily left many issues open. We did not discuss how the likelihood and convention databases are to be populated. Another open question is which aspects of the utterance are listed in the convention database: that is, do the conventions relate to lexical items, syntactic categories (such as VP), or some other form of regular expression? Finally, a very general question concerns the nature of the possible intentions themselves, an issue that has been explored from many directions. However, although we concede that the correct set of intentions must be posited in order to precisely simulate human behaviour, we would argue that the use of any plausible proxy set should be adequate in principle to achieve a close approximation to this behaviour.

In future work we aim to explore the capabilities of this model through a range of qualitative and quantitative tests. The model gives rise to testable predictions as to a wide range of behaviours. These include the attribution of communicative intentions, the planning of conversational turns, and instances of repair. We feel that the model has considerable practical potential in providing enhanced artificial discourse capabilities, and that if this promise is borne out, it could also have substantial implications for the modelling of dialogic behaviour in natural language.

References

Herbert H. Clark and Catherine Marshall. 1981. Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber and I. A. Sag (eds.), Elements of Discourse Understanding. New York: Cambridge University Press. 10-63.

H. Paul Grice. 1957. Meaning. Philosophical Review, 67: 377-388.

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

Claude E. Shannon. 1948. A mathematical theory of communication. Bell System Technical Journal, 27: 379-423.

Robert Stalnaker. 1978. Assertion. Syntax and Semantics, 9: 315-332.

Tanya Stivers et al.. 2009. Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences of the United States of America, 106: 10587-10592.

Spatial descriptions in discourse: choosing a perspective

Simon Dobnik
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Box 200, 405 30 Göteborg, Sweden
simon.dobnik@gu.se

The semantics of natural language spatial descriptions such as “to the left of” and “above” involve (i) perceptual knowledge obtained from scene geometry , (ii) world knowledge about the objects involved , and (iii) shared knowledge that is established as the common ground in discourse. Dialogue partners coordinate all three types of meaning when describing and interpreting visual scenes. One example of (iii) is the perspective or the reference frame (RF). For example, the table may be “to the left of the chair”, “to the right of the chair”, “behind the chair” or “South of the chair”. The perspective is determined by some point in the scene called the *viewpoint* (VPT). There are three ways in which the VPT is set in human languages (Levinson, 2003): (i) *relative RF:* by some third object distinct from the located and reference objects (the speaker, the hearer, the sofa); (ii) *intrinsic RF:* by the reference object itself (the chair); or (iii) *extrinsic RF:* by some global reference point (the North).[1] The geometric spatial template or potential fields are projected within the framework defined by the VPT (Maillat, 2003). The RF may be inferred from the perceptual context if given some configuration of the scene a spatial description is true only in one RF. It follows that when interpreting and generating spatial descriptions humans rely on verification of spatial templates in different RFs which requires considerable computational complexity (Steels and Loetzsch, 2009). Alternatively, it may be described linguistically “from your view” or “from there”.

(Watson et al., 2004) show experimentally that (i) participants are significantly more likely to use an intrinsic RF after their partner used an intrinsic RF, compared when the partner used a relative RF (with the speaker as the VPT); (ii) participants are significantly more likely to use intrinsic RF when the objects are aligned horizontally (their typical alignment in the world) than when they are aligned vertically; (iii) the alignment of the RFs is not due to the lexical priming caused by using the same preposition. (Andonova, 2010) shows for the map task that overall partners align with the primed route or survey perspective set by the confederate if priming is consistent – when the confederate changes the perspective only once in the middle of the session. On the other hand, if the confederate regularly alternates between the perspectives their partner has nothing to prime to. The self-assessed spatial ability (using a standardised test) is also important – low ability participants only align with the primed perspective when the switch is from the survey to the route perspective which is otherwise also the most frequently used one.

Our interest is to implement these and similar strategies as information state update rules in a dialogue manager such as GoDiS (Larsson, 2002). In such a model each conversational agent must keep a record of their own RF and that of their partner in the common ground. The RFs are updated following perceptual verification and an alignment strategy. The proposal is a move towards a more natural interpretation and generation of projective spatial descriptions in an artificial conversational agent compared to our previous attempt where the RF parameters were not specifically included in the model but some RF

[1] Sometimes (mostly for route descriptions) it is distinguished between speaker-oriented (egocentric) and external (allocentric) perspective or between (i) route and (ii) survey perspective. The model we follow is more specific.

knowledge has nonetheless been learned with machine learning. We proceed as follows:

1. Collect a corpus of dialogue interactions containing projective spatial descriptions made in a room scene.
2. Annotate the dialogue utterances with an XML annotation scheme which identifies perceptual states, objects in focus, utterances, turns, speakers, located objects, RFs, VPTs, spatial relations, ref. objects, etc.
3. Replicate the literature findings on the RF usage in our dataset.
4. Repeat the experiments from (1) but where one of the participants is a dialogue manager following an RF strategy. Allow humans conversational partners to rate the performance of the system.
   (a) Always use the relative RF to yourself.
   (b) Always align to the RF used by your partner in the previous turn.
   (c) For each turn select the RF randomly.
   (d) Keep a randomly chosen RF for n turns, then change.

   To prevent over-agreement with the system the evaluators should, ideally, compare pairs of strategies and select the preferred one.

We collect our data and later test the interaction in an online experimental environment specifically developed for this purpose (`http://goo.gl/8KLja`). Participants may create sessions to which they invite other participants and complete them interactively in their own time. During a session each participant sees a 3d generated image of a room containing some furniture. The image also contains two avatars: the one with their back towards the participant is the participant and the one facing the participant from the opposite side of the room is their partner. This is explained to the participants in the instructions and different representations are used to avoid the confusion. The other participant sees the room from the opposite side. The participants communicate via a text chat interface which allows unrestricted entry of text and also logs and partially annotates both the conversation and the perceptual information in the background.

By the time of writing this abstract we conducted two pilot studies for which we completed stages 1 to 3 of our plan. In the first pilot study (7 conversations) we used a room with four distinct entities (two participants, a chair and a sofa) arranged around a table in the middle which was placed on a carpet. We instructed the participants to talk about the location of the objects in the scene. Although this method was good in encouraging spontaneous conversations it had two shortcomings: (i) the participants produced less spatial descriptions than desired (11.9 per conversation) as they also discussed their opinions about the objects, etc.; and (ii) they spontaneously took on roles where one was asking questions and the other was giving answers and therefore the conversations included were very few cases of interaction that we were looking for. To overcome these difficulties we designed a second pilot study for which we (i) only used one kind of objects (the chairs), (ii) restricted the conversational interaction to pair of turns where in the first turn one participant describes which chair they chose (one is automatically selected for them and marked with an arrow) and then in the second turn the partner selects that chair on their view of the room. The roles are reversed in the next turn. Thus, we get a series of dialogue turns from which we record (i) speaker's strategy for RF choice; (ii) the hearer's understanding of the description. The latter is important as a particular description may be true under more than one RF.

References

Elena Andonova. 2010. Aligning spatial perspective in route descriptions. In Christoph Hölscher, et al., editors, *Spatial Cognition VII*, volume 6222 of *Lecture Notes in Computer Science*, pages 125–138. Springer Berlin, Heidelberg.

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

Stephen C Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*, volume 5. Cambridge University Press, Cambridge.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford, UK, May.

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.

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, August.

Modeling Referring Expressions with Bayesian Networks

Kotaro Funakoshi Mikio Nakano
Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako,
Saitama 351-0188, Japan
{funakoshi,nakano}@jp.honda-ri.com

Takenobu Tokunaga Ryu Iida
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro,
Tokyo 152-8550, Japan
{take,ryu-i}@cl.cs.titech.ac.jp

Abstract

A probabilistic approach to the resolution of referring expressions for task-oriented dialogue systems is introduced. The approach resolves descriptions (e.g., “the blue glass”), anaphora (e.g., “it”), and deixis (e.g., “this one” w/ pointing gesture) in a unified manner. In this approach, the notion of reference domains serves an important role to handle context-dependent attributes of entities and references to sets. Previously we reported the evaluation results in a puzzle solving task. This paper briefly explains the approach and discusses the issues in two work-in-progress application projects.

1 Introduction

Referring expressions (REs) can be classified into three categories: descriptions, anaphora, and deixis. Dialogue systems (DSs) are expected to handle all the three categories of REs.

We employ a Bayesian network (BN) to model a RE. One of the two major novelties of the approach is our probabilistic formulation that handles the above three kinds of REs in a unified manner. The other is bringing reference domains (RDs) (Salmon-Alt and Romary, 2001) into that formulation. RDs are sets of referents implicitly presupposed at each use of REs. By considering RDs, our approach can appropriately interpret context-dependent attributes. In addition, by treating a reference domain as a referent, REs referring to sets of entities are handled, too.

Our approach presupposes a certain amount of manual implementation of domain-dependent knowledge by developers. Therefore, it would not be suited to general text processing. However, it has the potential to be used for any task-oriented applications such as personal agents in smart phones, in-car systems, robots, etc.

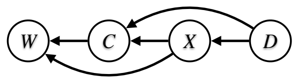


Figure 1: WCXD fundamental structure.

2 Bayesian Network-based Modeling

Each REBN (Referring Expression Bayesian Network) is tailored for a RE in the context at the moment. Its structure is determined by the syntactic and semantic information in the RE and probability tables are determined by the context. Here, we describe REBNs briefly. The details and an evaluation are found in (Funakoshi et al., 2012).

2.1 Structures

Figure 1 shows the fundamental network structure of REBNs. We call this structure WCXD. The four nodes (random variables) W, C, X, and D represent an observed word, the concept denoted by the word, the referent of the RE, and the presupposed RD, respectively. Here, a *word* means a lexical entry in the system dictionary. Each REBN is derived from the WCXD structure.

2.2 Domains of random variables

A REBN of N words referring to one entity has $2N + 2$ discrete random variables: $W_1, \ldots, W_N, C_1, \ldots, C_N, X$, and D. The domain of each variable depends on the corresponding RE and the context at the moment. Here, $\mathfrak{D}(V)$ denotes the domain of a variable V.

$\mathfrak{D}(W_i)$ contains the corresponding observed word w_i and a special symbol ω that represents other possibilities, i.e., $\mathfrak{D}(W_i) = \{w_i, \omega\}$. Each W_i has a corresponding node C_i.

$\mathfrak{D}(C_i)$ contains M concepts that can be expressed by w_i and a special concept Ω that represents other possibilities, i.e., $\mathfrak{D}(C_i) = \{c_i^1, \ldots, c_i^M, \Omega\}$. c_i^j ($j = 1 \ldots M$) are looked up from the system dictionary.

$\mathfrak{D}(D)$ contains $L+1$ RDs recognized up to that point in time, i.e., $\mathfrak{D}(D) = \{@_0, @_1, \ldots, @_L\}$. $@_0$ is the ground domain that contains all the individual entities to be referred to in a dialogue. At the beginning of the dialogue, $\mathfrak{D}(D) = \{@_0\}$. Other L RDs are incrementally added in the course of the dialogue.

$\mathfrak{D}(X)$ contains all the possible referents, i.e., K individual entities and $L + 1$ RDs. Thus, $\mathfrak{D}(X) = \{x_1, \ldots, x_K, @_0, \ldots, @_L\}$.

2.3 Probability tables

A REBN infers the referent (i.e., the true value of node X) using four types of probability tables.

Realization model: $P(W_i|C_i, X)$

$P(W_i = w|C_i = c, X = x)$ is the probability that a hearer observes w from c and x which the speaker intends to indicate.

Relevancy model: $P(C_i|X, D)$

$P(C_i = c|X = x, D = d)$ is the probability that concept c is chosen from $\mathfrak{D}(C_i)$ to indicate x in d. Developers can implement task domain semantics in $P(C_i|X, D)$. By considering d, context-dependent attributes are handled.

Referent prediction model: $P(X|D)$

$P(X = x|D = d)$ is the probability that entity x in RD d is referred to, which is estimated according to the contextual information (such as gaze) at the time the RE is uttered but irrespective of attributive information in the RE.

Domain prediction model: $P(D)$

$P(D = d)$ is the probability that d is presupposed at the time the RE is uttered, which is estimated according to the saliency of d.

3 Work-in-Progress Apps and Issues

Currently we are working on two different applications: *Map-search* as a mobile/PC application and *Object-fetch* as a robotic application. In Map-search, the user can search locations on a map and identify a location to query the information of the location or to get a navigation to the location. In Object-fetch, the user makes a robot identify an object in the user's home or office to fetch him/her it. By applying REBNs to these domains different from each other and from the Tangram task with which we made the first evaluation, we will be able to verify the quantitative performance and qualitative ability of our approach in diverse aspects. For example, in Map-search, the number of referents can be huge while Tangram has only 7 referents. Therefore, computational complexity will be an important issue for realtime operation. It is unrealistic to consider all locations every time. We will have to devise a way to efficiently limit the number of candidates for each time without excluding true referents.

Not limited to Object-fetch but especially in it, handling of unknown objects is vital, while all objects are known in Tangram. The robot must recognize a RE to an object that it does not know. For this purpose we can introduce χ for an unknown referent in $\mathfrak{D}(X)$. Hopefully, χ will have the highest probability for REs to unknown objects. Uncertainty due to speech recognition errors, unknown words, and unknown concepts is also a severe issue. There is a possibility that adjusting the parameter ϵ (here, $P(W = w|C = \Omega, X) = \epsilon$) eases the problem. The larger ϵ is, the more $P(X|D)$ influences inference results, i.e., the contextual information outside the RE gets more importance. For example, in a low signal-noise ratio environment, the robot could selectively rely on the context by increasing ϵ.

In both applications, spatial relations are important to identify referents. To handle relations, we are going to introduce another type of node for relations in REBNs to combine multiple REBNs into one.

System design methodology is the last but not least issue. While REBNs allow different design patterns of the world inherent in each application, the best design pattern seems to depend on each. For example, using the set of the location IDs in a database as $\mathfrak{D}(X)$ seems reasonable for Map-search. However, this design pattern does not work with Object-fetch. Object-fetch requires the object IDs in the robot's database to be included in $\mathfrak{D}(C)$. Through building Map-search and Object-fetch in parallel, we would like to clarify different design patterns and the conditions to chose a design pattern for each application.

References

K Funakoshi, M. Nakano, T. Tokunaga, and R. Iida. 2012. A unified probabilistic approach to referring expressions. In *Proc. SIGDIAL 2012*.

S. Salmon-Alt and L. Romary. 2001. Reference resolution within the framework of cognitive grammar. In *Proc. Intl. Colloqium on Cognitive Science*.

Helping the medicine go down: Repair and adherence in patient-clinician dialogues

Christine Howes, Matt Purver, Rose McCabe, Patrick G.T. Healey and Mary Lavelle
Queen Mary University of London
Interaction, Media and Communication Research Group
c.howes@qmul.ac.uk

Abstract

Repair is crucial in achieving and maintaining shared understanding in dialogue. Recent work on consultations between patients with schizophrenia and psychiatrists has shown that adherence to treatment correlates with patterns of repair. We show that distributions of repair in consultation dialogues are different to those in general conversation.We investigate whether particular types of repair can be detected from high-level dialogue features and/or lexical content, with encouraging results. We further explore whether we can predict adherence directly from these features. The results indicate that prediction appears to be possible from low-level lexical content.

1 Introduction

How conversational partners achieve and maintain shared understanding is crucial in the understanding of dialogue. One such mechanism, repair, is pervasive and highly systematic. In Schegloff et al. (1977), repairs are described in terms of who initiates the repair, who completes it, and in what position it is completed.

A speaker can repair their own utterance in the course of producing it – a *position 1 self-initiated self-repair* (P1SISR), by repeating (articulation), reformulating (formulation), or adding something (transition space). They may also repair one of their own utterances following someone else's – a *position 3 self-initiated self-repair* (P3SISR). A speaker can also repair another's utterance – a *position 2 other initiated other repair* (P2OIOR) or signal misunderstanding – a *position 2 next turn repair initiator* (P2NTRI) prompting the original speaker to repair their prior utterance – a *position 3 other initiated self repair* (P3OISR). See table 1 for examples.

Type		Example
P1SISR(A)	Dr:	You probably have seen so many psychiatrists *o- o-* **over** the years
P1SISR(F)	Dr:	*Did you feel that* **did you despair so much that** you wondered if you could carry on?
P1SISR(TS)	P:	Where I go to do some *printing*. **Lino printing**
P3SISR	Dr:	*Clozaril* or
	P:	Yeah
	Dr:	**Clozapine** yes
P2OIOR	Dr:	rather than *the diazepam* which I don't think ... is going to do you any good
	P:	**The valium**
P2NTRI	Dr:	*It doesn't happen in real life does it?*
and	P:	What do you mean by real life?
P3OISR	Dr:	**you can't- there are no messages coming from the television to people are there?**

Table 1: Repair types (repair bold; repaired italics)

McCabe et al. (in preparation) analysed repair in dialogues between patients with schizophrenia and their psychiatrists. More patient led clarification, e.g. clarifying the psychiatrist's utterance with P2NTRIs, was associated with better treatment adherence 6 months later. Explaining the link between communicative patterns and adherence has both clinical and theoretical implications.

2 Repair in different dialogue contexts

We compared the repair data from Colman and Healey (2011)[1] with that from McCabe et al. (in preparation). These were annotated for instances of repair using the same protocol.

As shown in figure 1, although all types of dialogue exhibit the preference for self repair (Schegloff et al., 1977), this is especially the case in the clinical dialogues. Conversely, in the clinical dialogues there are fewer P2NTRIs and P3OISRs.

[1]This study looked at the demographic portion of the British National Corpus, and HCRC Map Task dialogues.

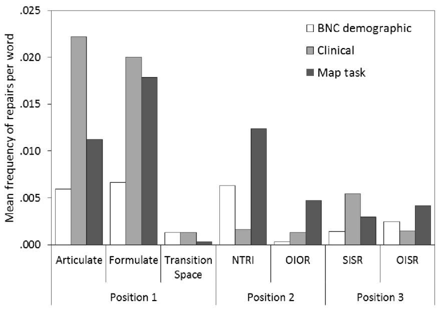


Figure 1: Repair per word by dialogue context

3 Classification Experiments

We first investigate the automatic detection of P2NTRIs, and then prediction of adherence directly, using the Weka machine learning toolkit (Hall et al., 2009) and the support vector machine implementation SVMLight (Joachims, 1999).

3.1 Detecting Repair

We defined a set of turn-level features (table 2) extracted automatically and likely to correlate with P2NTRIs. Words used by the patient were used to extract (optional) lexical unigram features.

Measure	Description
Speaker	Doctor, Patient, Other
NumWords	Number of words in turn
OpenClassRepair	Contains *pardon, huh* etc
WhWords	Num of wh-words (*what, who, when*)
Backchannel	Num of backchannels (*uh-huh, yeah*)
FillerWords	Number of fillers (*er, um*)
RepeatedWords	Words repeated from preceding turn
MarkedPauses	Number of pauses transcribed
OverlapAny	Number of portions of overlapping talk
OverlapAll	Entirely overlapping another turn

Table 2: Turn-level features

The classification task is to categorise each patient turn as containing a P2NTRI or not. The target class is very sparse: 170 of 20,911 turns were P2NTRIs, so a weighted SVM cost function was used. Performance was evaluated using 5-fold cross-validation. As shown in Table 3, absolute F-scores are low due to target class sparsity.

Target	Features	F (%)	P (%)	R (%)
P2NTRI	OCRProportion	35.8	85.7	22.6
P2NTRI	High-level	41.4	42.8	40.6
P2NTRI	All	44.0	44.9	43.6

Table 3: Repair detection

3.2 Predicting Adherence

We now turn to classifying each dialogue according to the level of adherence after 6 months. The features used were similar to those in the turn-level experiments, calculated over the dialogue.

Given the small size of the dataset (77 instances) and large possible feature space when using lexical features, we allowed only words mentioned >40 times, and selected the most predictive 10-20 features based only on the training set in each fold of the cross-validation.

As Table 4 shows, the performance using best selected features is good; however, all features selected are unigram lexical features. High-level features do not prove useful.

Features	F (%)	P (%)	R (%)
High-level	35.5	27.0	51.9
Best features	70.3	70.3	70.3

Table 4: Adherence prediction

4 Discussion

Patient led clarification is rare, leading to a highly unbalanced dataset. Although P2NTRIs can be predicted, the sparsity of the data mean they are not sufficient to predict adherence. Patient led clarification is not straightforwardly associated with any high-level, general dialogue factors to allow us to accurately classify the adherent patients.

However, there is a link between patients' conversational behaviour and their subsequent adherence to treatment, as seen in the results of experiments using words as features. Further work is needed to clarify what this link is and whether we can come up with a usable metric for predicting probable adherence from dialogue transcripts.

References

M. Colman and P. G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of CogSci*, pages 1563–1568, Boston, MA.

M. Hall, E. Frank, G. Holmes, et al. 2009. The WEKA data mining software: An update. *SIGDKDD Explorations*, 11(1):10–18.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press.

R. McCabe, M. Lavelle, S. Bremner, et al. in preparation. Shared understanding in psychiatrist-patient communication: Association with treatment adherence in schizophrenia.

E.A. Schegloff, G. Jefferson, and H. Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

A spoken dialogue interface for pedestrian city exploration: integrating navigation, visibility, and Question-Answering

Srinivasan Janarthanam[1], Oliver Lemon[1], Xingkun Liu[1], Phil Bartie[2], William Mackaness[2], Tiphaine Dalmas[3] and Jana Goetze[4]
[1]Interaction Lab, Heriot-Watt University, Edinburgh
[2] School of GeoSciences, University of Edinburgh
[3]School of Informatics, University of Edinburgh
[4]KTH Royal Institute of Technology, Stockholm, Sweden
sc445,o.lemon,x.liu@hw.ac.uk, philbartie@gmail.com,
william.mackaness@ed.ac.uk, tiphaine.dalmas@aethys.com, jagoetze@kth.se

Abstract

We demonstrate a spoken dialogue-based information system for pedestrians. The system is novel in combining geographic information system (GIS) modules such as a visibility engine with a question-answering (QA) system, integrated within a dialogue system architecture. Users of the demonstration system can use a web-based version (simulating pedestrian movement using StreetView) to engage in a variety of interleaved navigation and QA conversations.

1 Motivation

Although navigation and local information are available to users through smartphone apps, there are still important problems such as how such information is delivered safely and proactively, and without cognitively overloading the user. (Kray et al., 2003) suggested that the cognitive load of information presented in textual and speech-based interfaces is medium and low respectively when compared to more complicated visual interfaces. Another important challenge is to bring different sources of data together and present information appropriately based on the dialogue context. Our objective, therefore, is to build a hands-free and eyes-free system that engages the pedestrian user by presenting all information and receiving user inputs through speech only. The system integrates a City Model and a Visibility Engine to identify visible landmarks, a Pedestrian Tracker to improve the GPS positioning of the user, and a question-answering (QA) system to enable users to explore information about the city more freely than with a graphical interface (see example in table 1).

User: Take me to Princes Street.
System: Turn left on to South Bridge and walk towards the tower in front of you.
...
System: Near you is the famous statue of David Hume.
User: Tell me more about David Hume.
System: David Hume was a Scottish philosopher....

Table 1: An example interaction with the system

2 Architecture

The architecture of the current system is shown in figure 1.

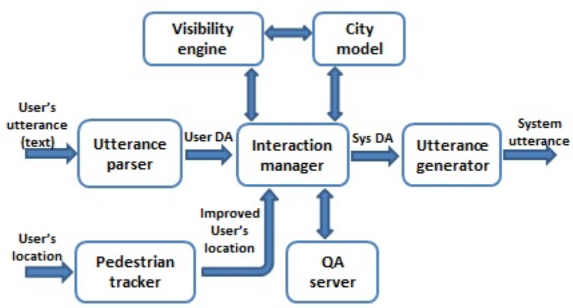


Figure 1: System Architecture

2.1 Dialogue interface

The dialogue interface consists of an utterance parser, an Interaction Manager and an utterance generator. The Interaction Manager (IM) is the central component of the system, which provides the user with timely navigational instructions and interesting PoI information. It receives the user's input in the form of a dialogue act (DA) and the user's location in the form of latitude and longitude. Based on these inputs and the dialogue context, it estimates the user's orientation, and it responds with system output dialogue act, based on a dialogue policy. The utterance generator is a natural language generation module that translates the system's DAs into surface text, using the Open CCG toolkit (White et al., 2007).

2.2 Pedestrian Tracker

Using Global Navigation Satellite Systems (GNSS) (e.g. GPS, GLONASS) this module provides user positioning information. Since urban environments can be challenging with limited sky views, and hence limited line of sight to satellites, this module improves on the reported user position by combining smartphone sensor data (e.g. accelerometer) with map matching techniques, to determine the most likely location of the pedestrian (Bartie and Mackaness, 2012).

2.3 City Model

The City Model is a spatial database containing information about thousands of entities in the city of Edinburgh. These data have been collected from a variety of existing resources such as Ordnance Survey, OpenStreetMap and the Gazetteer for Scotland. It includes the location, use class, name, and street address of many entities. The model also includes a pedestrian network (streets, pavements, etc) which can be used to calculate routes for the user.

2.4 Visibility Engine

This module identifies the entities that are visible to the user using a 2.5D representation of the city. This information is used by the IM to generate effective navigation instructions. E.g. "Walk towards the castle", "Can you see the tower in front of you?", "Turn left after the large building on your left, after the junction" and so on.

2.5 Question-Answering server

The QA server currently answers a range of *definition* questions. E.g., "Tell me more about the Scottish Parliament", "Who was David Hume?", etc. QA identifies the entity focused on in the question using machine-learning techniques (Mikhailian et al., 2009), and then proceeds to a textual search on texts from the Gazetteer of Scotland and Wikipedia, and definitions from WordNet glosses.

3 Web-based User interface

For the purposes of this (necessarily non-mobile) demonstration, we present a web-based interface that simulates users walking in a 3D city environment. Users will be able to provide speech or text input. The web-based client is a JavaScript/HTML program running on the user's web browser. For a detailed description of this component, please refer to (Janarthanam et al., 2012). A simulated real world is presented to the user visually using a Google Streetview client (Google Maps API). It allows the user to simulate walking around in real streets using arrow keys. The user can interact with the dialogue system using speech or text, which is sent to the system along with the user's location. The system's utterances are synthesized using the Cereproc text-to-speech engine.

Acknowledgments

The research has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270019 (SPACEBOOK project www.spacebookproject.org).

References

P. Bartie and W. Mackaness. 2012. D3.4 pedestrian position tracker. Technical report, The SPACEBOOK Project (FP7/2011-2014 grant agreement no. 270019).

S. Janarthanam, O. Lemon, and X. Liu. 2012. A web-based evaluation framework for spatial instruction-giving systems. In *Proc. of ACL 2012, South Korea*.

C. Kray, K. Laakso, C. Elting, and V. Coors. 2003. Presenting route instructions on mobile devices. In *Proceedings of IUI 03, Florida*.

A. Mikhailian, T. Dalmas, and R. Pinchuk. 2009. Learning foci for question answering over topic maps. In *Proceedings of ACL 2009*.

M. White, R. Rajkumar, and S. Martin. 2007. Towards Broad Coverage Surface Realization with CCG. In *Proc. of the UCNLG+MT workshop*.

Influencing Reasoning in Interaction: a Model

Haldur Õim
University of Tartu
J. Liivi 2
Tartu, 50409 Estonia
haldur.oim@ut.ee

Mare Koit
University of Tartu
J. Liivi 2
Tartu, 50409 Estonia
mare.koit@ut.ee

Abstract

We are modeling human dialogues where one participant tries to influence the reasoning process of the other in order to get him to perform an action. Our aim is to build a dialogue system which would interact with a user in a `natural human way´. In our model, reasoning is presented as a process of evaluating different aspects of the action. To describe the influencing of reasoning, we introduce two conceptual structures: communicative strategies and communicative tactics.

1 Aims and Background

One of the central tasks of pragmatics is to explain the mechanisms by which people reach their goals in communication; to this end, also several computer models have been built (Jokinen, 2009; Ginzburg and Fernández, 2010). We believe that the central task is here to explain the process we call `influencing the reasoning of a communication partner´. A computer system should include a model of reasoning and account for the means used by people to influence the reasoning of others. This view has got strong support from the (evolutionary) psychologists who claim that the original function of the human reasoning is argumentative: to devise and evaluate arguments intended to lead partners to make/accept certain decisions (see Mercier and Sperber, 2011).

In our *Interaction Model* (IM) we treat dialogues where one of the participants (*A*) is trying to achieve the partner´s (*B*) decision to perform an action (*D*), and have worked out a corresponding computer model. We follow the general ideas of the BDI model (Allen, 1995) and have elaborated it in the aspects relevant for us. The central submodels of IM are: (1) *Model of reasoning subject* (RS) which contains *Model of motivational system* (MS) and *Reasoning model* (RM), by which the process of evaluating (weighing) the relevant aspects of *D* is carried out; and (2) *Models of communicative strategies* (ComStr) *and communicative tactics* (ComT), by which the process of *Influencing the reasoning* is treated (Koit and Õim, 2004). The empirical data of our study are taken from the Estonian dialogue corpus[1].

2 Model of Reasoning Subject

2.1 Motivation and Reasoning

We assume that the reasoning process concerning *D* is *triggered* by one of three *motivational factors* of RS: (1) RS may like to perform *D* (*wish*-factor), (2) RS may assume that *D* is useful for reaching some goal (*needed*-factor), or (3) *D* is obligatory (*must*-factor). Together, these factors constitute the MS of the reasoning subject.

MS is used by RS in reasoning about *D*, by weighing the positive/negative aspects of *D* departing from these factors: *pleasant/unpleasant, useful/harmful, obligatory/prohibited.* If the positive aspects (pleasant, etc.) weigh more, RS will decide to do *D*, otherwise the decision will be not to do *D*.

[1] http://www.cs.ut.ee/~koit/Dialoog/EDiC.html

Thus, we assume that RS is able to `sum up´ the results of weighing. In our model this assumption is formally realized so that the evaluated aspects are represented as *scales* that take numerical values, or *weights.* The scales are different. For example, the weights of pleasant/unpleasant scales, *w(pleasant)*, *w(unpleasant)* have values from 0 to n, where n depends on the participant, whereas weights of obligatory and prohibited scales have values 0 or 1. At the same time, the *punishment* scale is connected with these scales, the weights of *w(punishment-not-D)* and *w(punishment-D)*, correspondingly, can again have scalar numerical values. The scales are not independent: what is useful can also be pleasant; punishment is unpleasant, etc. One more motivational aspect is *Resources* (mental, physical, etc.) needed to carry out *D*.

Reasoning procedures (RP) are represented as algorithms of going through the weights of relevant aspects of the action depending on the initiating factor (*Wish, Needed,* or *Must*). Algorithms are represented as decision trees including *yes-no* questions. For instance, in case of RP triggered by *Wish*-factor one question is: *Is w(pleasant) > w(unpleasant) + w(harmful)?* There are three RPs in our model: *Wish, Needed,* and *Must*. The first step of all the procedures is: *Are there enough Resources for doing D? If not, then do not do D,* and every path of a tree ends with a decision: *Do D* or *Do not do D.*

2.2 Influencing Reasoning

If after the *A*´s first turn (request, proposal, etc.) *B* does not agree to do *D* (and *A* does not give up), interaction follows: *A* tries to influence the reasoning of *B*, departing, according to our approach, from the MS and RM of *B*. The influencing consists in manipulating the weights of the relevant scales on MS/RS of *B*, information about which *A* gets from *B*´s (counter-)arguments during the interaction. This `manipulation´ presupposes certain reasoning procedures of *A*, `reasoning about reasoning (of *B*)´, as the output of which he will choose a coherent line of action: which weights in RS of *B* to increase or downgrade. Here we distinguish between two levels of procedures: *communicative tactics* and *communicative strategies*. ComT-s are procedures determined by the choice of the primary motivational factor (*Wish, Needed,* or *Must*). Accordingly, we have three ComT-s in our model which we call *Enticing, Persuading,* and *Threatening*. They consist in increasing the weights of *w(pleasant)*, *w(useful)*, *w(obligatory)*, correspondingly, while downgrading the negative weights relevant for *B*. For instance, if *A* has chosen ComT *Enticement* but *B* points at harmful consequences of *D*, then *A* tries to downgrade *w(harmful)* for *B*. ComStr-s are higher order procedures that regulate the possible choices between ComT-s in a certain interaction. Concretely, in our model two kinds of ComStr-s are important: *Attack* and *Defense* (these apply to *A* as well as to *B*). In the first case, the participant tries to press his goal on the partner, in the second, he averts taking over the partner´s goal. The choice between these ComStr-s clearly restricts the use of possible ComT-s.

3 Future Work

We will include contextual dimensions to our reasoning-in-interaction model, first, involving personal background of the participants: their social relationships (status, distance: friends-adversaries); and second, characteristics of the interaction (rueful, vehement, etc.).

Acknowledgments

This work is supported by the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS) and the Estonian Science Foundation (grant 9124).

References

James Allen. 1995. *Natural Language Understanding.* 2nd ed. The Benjamin/Cummings Publ. Comp., Inc.

J. Ginzburg and R. Fernández. 2010. Computational Models of Dialogue. *The Handbook of Computational Linguistics and Natural Language Processing*, Clark, Fox, and Lappin (eds.). Wiley Blackwell Publishers, 429-481.

Kristiina Jokinen. 2009. *Constructive Dialogue Modelling: Speech Interaction and Rational Agents.* John Wiley & Sons Ltd.

Mare Koit and Haldur Õim. 2004. Argumentation in the Agreement Negotiation Process: A Model that Involves Natural Reasoning. *Proc. of the Workshop on Computational Models of Natural Argument. ECAI,* Valencia, Spain, 53-56.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34, 57-111.

Rhetorical Structure for Natural Language Generation in Dialogue

Amy Isard and **Colin Matheson**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh, United Kingdom
{amy.isard,colin.matheson}@ed.ac.uk

Abstract

Many traditional dialogue systems use simple predicates to send information between a Dialogue Manager and a Natural Language Generation system. We propose a flexible RST-style interface to allow for more complex structures and multimodal output, and we place the first stage of content planning under the control of the dialogue management system with access to a system-wide information state.

1 Introduction

In many dialogue systems, a Dialogue Manager (DM) sits at the centre of the interaction, taking user inputs and sending an output specification to a Natural Language Generation (NLG) system. In multimodal systems, there may be a Dialogue and Interaction Manager, and the outputs may go via an intermediate stage where the different modalities are synchronised, and in other cases a planner may take over the duties of dialogue management, but there is usually a need to store the history and to specify the language output. Where the system has a physical component, facial and hand gestures may also be specified.

We propose a flexible interface between the DM and the output modalities which can be used in a variety of human-machine dialogue domains. It is based on RST structures (Mann and Thompson, 1988), and is related to work by (Stent et al., 2004) who used a similar approach in MATCH, their text-based restaurant recommendation system. However in MATCH, the DM sends high-level goals to the text planner, whereas in our systems, the DM performs part of the content selection task (in some cases by communicating with a domain planner and/or knowledge representation module), and sends a structured representation to the NLG system. The potential content to express can come from a wide range of sources, including the dialogue history, the domain knowledge or the task plan. For example, the task plan may describe sequences of actions which need to be carried out in constructing objects (Foster and Matheson, 2008).

This work can be seen in the broader context of attempting to integrate language processing, dialogue management, and NLG more closely.

2 Use Cases

To date, we have used our RST representation in three working systems in varied domains:

- the JAST system (Foster and Matheson, 2008; Foster et al., 2009), which allows a human to collaborate with a robot in building simple wooden toys
- the JAMES system (Petrick and Foster, 2012; Petrick et al., 2012), a robot bartender
- the Beetle II system (Dzikovska et al., 2011), a tutorial dialogue system for basic electricity and electronics

The top level rhetorical structures which we use are the following:

enablement where one situation or action is necessary (but not always sufficient) for another situation to action to occur. e.g. "to build a tower, insert the green bolt through the red cube and screw it into the blue cube" (JAST).

elaboration where one piece of content adds further information about an object which has already been mentioned e.g. "the battery in circuit 5 is in a closed path which does not contain a bulb" (Beetle II).

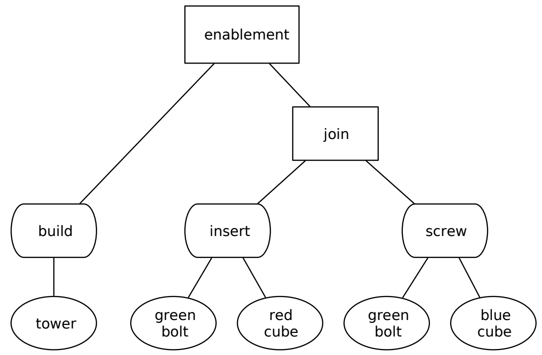


Figure 1: Graph representation of enablement and join relations

```
<output>
 <objects>
  <obj id="o1" type="bolt" color="green"/>
  <obj id="o2" type="cube" color="red"/>
  <obj id="o3" type="cube" color="blue"/>
  <obj id="o4" type="tower"/>
 </objects>
 <rst>
  <relation type="enablement">
   <pred action="build" result="o4"/>
    <relation type="join">
     <pred action="insert">
      <obj idref="o1"/>
      <obj idref="o2"/>
     </pred>
     <pred action="screw">
      <obj idref="o1"/>
      <obj idref="o3"/>
     </pred>
    </relation>
   </pred>
  </relation
 </rst>
 <actions>
  <action type="handover">
   <obj idref="o1"/>
  </action>
 </actions>
</output>
```

Figure 2: Multimodal RST XML for enablement relation

definition where one piece of information defines another e.g. “it means that the battery is damaged” (Beetle II).

join which signifies a simple aggregation of two pieces of content e.g. “hello, what would you like to drink” (JAMES).

A graph of an enablement relation from JAST is shown in figure 1, and a possible surface realisation for this is “to build a tower, insert the green bolt through the red cube and screw it into the blue cube”. The multimodal XML representation of the RST is shown in figure 2; as well as giving the content to be spoken, this specifies that the robot should hand over object o1 (the green bolt) to the user.

3 Conclusions

We designed a flexible interface for communications between the Dialogue Manager and the Natural Language Generation components in a dialogue system. We have used the interface in a number of different systems, and shown that it encourages the integration of multimodal output modalities. The systems we have described all use rule-based Dialogue Management or Planning, but the RST could also be used in a statistical dialogue system as long as the NLG component is grammar-based.

References

Myroslava O. Dzikovska, Amy Isard, Peter Bell, Johanna D. Moore, Natalie Steinhauser, and Gwendolyn Campbell. 2011. Beetle ii: an adaptable tutorial dialogue system. In *Proceedings of the 12th Annual SIGdial Meeting on Discourse and Dialogue, demo session.*

Mary Ellen Foster and Colin Matheson. 2008. Following assembly plans in cooperative, task-based human-robot dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial 2008)*, London, June.

Mary Ellen Foster, Manuel Giuliani, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. 2009. Evaluating Description and Reference Strategies in a Cooperative Human-Robot Dialogue System. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, July.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Ronald P. A. Petrick and Mary Ellen Foster. 2012. What would you like to drink? recognising and planning with social states in a robot bartender domain. In *International Workshop on Cognitive Robotics (CogRob 2012) at AAAI 2012*, Toronto, ON, Canada, July.

Ronald P. A. Petrick, Mary Ellen Foster, and Amy Isard. 2012. Social state recognition and knowledge-level planning for human-robot interaction in a bartender domain. In *AAAI 2012 Workshop on Grounding Language for Physical Systems*, Toronto, ON, Canada, July.

Amanda Stent, R. Prasad, and M. Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of ACL.*

Two semantical conditions for superlative quantifiers

Maria Spychalska
Ruhr-University Bochum
Department of Philosophy II
maria.spychalska@rub.de

Abstract

We discuss semantics of superlative quantifiers *at most n* and *at least n*. We argue that the meaning of a quantifier is a pair specifying a verification and a falsification condition for sentences with this quantifier. We further propose that the verification condition of superlative quantifiers has a disjunctive form, which should be interpreted in an epistemic way, that is as a conjunctive list of possibilities. We also present results of a reasoning experiment in which we analyze the acceptance rate of different inferences with superlative and comparative quantifiers in German. We discuss the results in the light of our proposal.

There is an on going debate concerning the right semantical interpretation of so-called superlative quantifiers, such as *at most n* and *at least n*, where n represents a bare numeral, e.g *two*. (Look inter alia: (Geurts & Nouwen, 2007), (Koster-Moeller et al, 2008), (Geurts et al., 2010), (Cummins & Katsos, 2010), (Nouwen, 2010), (Cohen & Krifka, 2011)). Generalized Quantifier Theory defines superlative quantifiers as equivalent to respective comparative quantifiers: *fewer than n* and *more than n*, that is:

$$at\,most\,n\,(A,B) \iff fewer\,than\,n+1(A,B) \quad (1)$$

$$at\,least\,n\,(A,B) \iff more\,than\,n-1(A,B) \quad (2)$$

It has been observed that in natural language those equivalences 1 and 2 might not hold, or at least they might not be accepted by language user based on pragmatical grounds. There are numerous differences between comparative and superlative quantifiers involving their linguistic use Geurts & Nouwen (2007), acquisition (Musolino, 2004), (Geurts et al., 2010), their processing (Koster-Moeller et al, 2008),(Geurts et al., 2010), as well as – what is our main focus – the inference patterns in which they occur. It has been for instance shown that majority of responders usually reject inferences from *at most n* to *at most n+*[1], although they accept the equivalent inference with comparative quantifiers (Geurts et al., 2010), (Cummins & Katsos, 2010).

In this paper we provide an algorithmic interpretation of superlative quantifiers that tries to explain the observed reasoning data. Furthermore we report on the results of a reasoning experiment that support our theoretical proposal.

We propose that the meaning of a quantifier as a pair $\langle C_F, C_V \rangle$, where C_V is a *verification condition* (specifies how to verify sentences with this quantifier) and C_F is a *falsification condition* (specifies how to falsify sentences with this quantifier). Verification and falsification conditions are to be understood *algorithmically* (as partial algorithms), with the "else" part of the *conditional instruction* being empty - thus, they verify (or falsify) the formulas only if their conditional test is satisfied. From a perspective of classical logic, these conditions should be dual, namely if C is a C_V condition for sentence ϕ, then C is a C_F condition for sentence $\neg\phi$, and vice versa. We further, however, observe that in the case of superlative quantifiers, there is a split between these two conditions. We suggest, that this bifurcation is a result of a pragmatic focus on the expressed borderline n.

[1] Here and below $n+$ denotes any natural number greater than n, while $n-$ denotes any natural number smaller than n.

1 Two semantic conditions for "at most"

Krifka (1999) points out that semantic interpretation is usually a pair that specifies when the sentence is true and when it is false. However, as he observes, a sentence *at most n x:* $\phi(x)$ says only that *more than n x:* $\phi(x)$ is false, and leaves a truth condition underspecified. In other words, the meaning of *at most n* provides an algorithm for falsifying sentences with this quantifier, but not (immediately) for verifying them. Consequently, the primal semantical condition of *at most n x:* $\phi(x)$ could be understood as an algorithm: "falsify when the number of x that are ϕ exceeds n", and would constitute what we understand by the falsification condition.

Definition 1 *(falsification condition for at most)*

$C_F(at\, most\, x : \phi(x)) := If\ \exists^{>n}x(\phi(x)),\ then\ falsify$

But how can we know when it is true? From the point of view of an algorithm, it is the "else" part of the conditional that should define the truth-condition. However a negation of a falsification condition is in sense *informationally empty*: it does not describe any concrete situation in which the given sentence can be verified. As a result, in those contexts that require to directly verify a sentence, we refer to a verification condition, which is specified independently. As expressing a positive condition, *at most n* may be understood as a disjunction *n or fewer than n* ("disjunctive *at most*").

$$\neg\exists^{>n}x\phi(x) \iff \exists^{<n+1}x\phi(x) \iff \exists^{=n}x\phi(x)\vee\exists^{<n}x\phi(x) \quad (3)$$

In order to define the verification condition, we adopt, following Zimmermann (2000), the view that disjunctive sentences in natural language are likely to get so-called epistemic reading that is they are interpreted as *conjunctive lists of epistemic possibilities*. According to the proposed solution a disjunction P_1 *or...or* P_n is interpreted as an answer to a question: *Q: What might be the case?* and, thus, is paraphrased as a (closed) list L: P_1 (might be the case) [and]... P_n (might be the case) (and nothing else might be the case). This results in the following reading of a disjunctive sentence:

Definition 2 *(Zimmermann, 2000)* $P_1\vee...\vee P_n \iff \diamond P_1\wedge...\wedge\diamond P_n\ and\ (closure)\ \forall P[\diamond P \rightarrow [P \between P_1\vee...\vee P \between P_n]]$

If we assume that disjunctions in natural language are likely to be interpreted as conjunctions of epistemic possibilities, then we get the following verification condition for *at most*:

Definition 3 *(epistemic interpretation of the verification condition for at most)*

$C_V^E(at\, most\, n\, x : \phi(x)) := If\ (\diamond\exists^{=n}x\phi(x) \wedge \diamond\exists^{<n}x\phi(x)), then\ verify$

$[and\ (closure)\ If\ \diamond\exists^{>n}x\phi(x),\ then\ falsify]$

The important point is the optional character of the closure. This bases on our assumption that the falsification and verification conditions are in a sense independent and only as a pair constitute the full semantic interpretation. Since the falsification condition, as defined in 3, is sufficient to account for the right semantical criterion of when the sentence with *at most n* is false, the closure of the verification condition is redundant and might or might not be considered in the reasoning process. The optional character of closure turns out crucial in evaluating validity of inferences with *at most n*.

It is easy now to observe that from $\diamond\exists^{=n}\phi(x)\wedge\diamond\exists^{<n}x\phi(x)$ one cannot infer $\diamond\exists^{=n+1}x\phi(x) \wedge \diamond\exists^{<n+1}x\phi(x)$: the conjunct $\diamond\exists^{=n+1}x : \phi(x)$ cannot be proven based on the premise, though it can be excluded only if the closure of the premise is applied. On the other hand, the inference: *n or fewer than n → n-1 or fewer than n-1* (in the epistemic interpretation) is blocked only due to closure of the conclusion. That is: $\diamond\exists^{=n}$ implied by the premise is contradicted by the closure of the conclusion, i.e. $\neg\diamond\exists^{>n-1}$. However, without the closure the implication holds (if the epistemic reading of the verification condition is applied).

2 "At least" and bare numerals

As an upward monotone quantifier, *at least n* appears to provide a clear verification algorithm: "verify when n x (that are ϕ) are found". Such a semantical interpretation would not, however, account for the linguistical differences between *at least n* and *more than n-1*.

Let us start with defining a falsification condition for *at least n* as follows:

Definition 4 *(falsification condition for at least)*

$C_F(at\ least\ n : x\phi(x)) := If\ \exists^{<n}x\phi(x),\ then\ falsify$

Defining a verification condition for "at least n" we first take into account following pragmatic focus that is put on the borderline n, which leads us to the disjunctive form of this quantifier: *(exactly)* *n or more than n*. Finally, we apply Zimmerman's epistemic interpretation.

Definition 5 (epistemic interpretation of the verification condition for *at least n*)
$C_F(at\ least\ n : x\phi(x)) := If\ (\diamond\exists^{=n}x\phi(x) \wedge \diamond\exists^{>n}x\phi(x)),\ then\ verify$

$and(closure)\ If\ (\bigvee_{i=0}^{n-1}\diamond\exists^{=i}x\phi(x)),\ then\ falsify$

Let us now show how the interpretation of the bare numeral n interacts with the validity of inferences (*n or more than n*)→ (*n-1 or more than n-1*), given the epistemic interpretation of disjunction. A bare numeral n (e.g. "two") can be interpreted as denoting any set of *at least* n elements, or a set of *exactly* n elements. Suppose now that n is interpreted with a closure: *exactly* n. It is easy to observe that, in such a case, *possible that n and possible that more than n* does not imply *possible that n-1 or possible that more than n-1*. The premise which is interpreted as in Definition 5 does not imply $\diamond\exists^{=n-1} \wedge \diamond\exists^{>n-1}$ (with closure $\bigwedge_{i=0}^{n-2} \neg\diamond\exists^{=i}x\phi(x)$) While $\diamond\exists^{>n-1}$ follows from both $\diamond\exists^{>n}$ and $\diamond\exists^{=n}$, the problematic element is $\diamond\exists^{=n-1}$, which is directly contradicted by the closure of the premise. But suppose that n does not get the "exact" reading, but it is interpreted barely as *there are n*. Then from *possible that n* we can infer *possible that n-1*, since the latter does not exclude the possibility that there is a bigger set of elements.

3 Main findings

In our pilot experiment on reasoning conducted on German native speakers: nearly 100% of responders accepted inferences from *at most n* to *not more than n* and vice versa, as well as from *n or fewer than n* to *at most n* (and vice versa), which suggests that they do see those expressions as equivalent. (Similarly for mutual inferences between: *at least n* and *not fewer than n*, and *at least n* and *n or more than n*). The inferences: *at least n*→ *at least n-* were accepted in only ca. 75% of cases, which suggests some, at least pragmatic mechanism, suppressing this inference. We propose that this rejection bases on the "exact" reading of bare numerals. It is worth to note that of subjects accepted inferences that base on the "at least" reading of bare numerals in almost 60% of cases, which highly correlated with their acceptance of the inferences: *n or more than n* → *n- or more than n-* ($p = .026$) and with their acceptance of inferences: *at least n* → *at least n-* ($p = .019$).

Furthermore, while inferences from *at most n* to *at most n+* were accepted only by 14% of responders, inferences from *not more than n* to *not more than n+* were already accepted by almost 32%. Thus, it seems that paraphrasing *at most n* to the negative form: *not more than n* facilitates the inference.

The results for the inferences with disjunctive forms of superlative quantifiers (*n or fewer than n* and *n or more than n*) are especially interesting. While logically valid inferences (*n or more than n*) → (*n- or more than n-*) are accepted by 65% people, the invalid inferences: (*n or more than n*) → (*n+ or more than n+*) are rather rejected (only 18% accept). The opposite effect, however, we get for disjunctive form of *at most*. The logically valid inferences (*n or fewer than n*) → (*n+ or fewer than n+*) are rather rejected (only 16% accept), whilst invalid inferences (*n or fewer than n*) → (*n- or fewer than n-*) are accepted in 39% of cases. The surprising result that subjects accepted the invalid inferences with "disjunctive at most" more frequently than the valid ones can be explained by our proposal. As we have proposed above, closure in the verification condition is optional, since the falsification condition is sufficient to account for the right semantics. However, if context enforces applying one of the semantical conditions (verification or falsification), then the other one might be ignored. While, from the perspective of classical logic it should be enough to use only one of the conditions (since the other can be defined via the first one), in the case of superlative quantifiers the epistemic reading of the verification condition creates the bifurcation in the meaning. This results in different inferential patterns in which those quantifiers occur, depending on what the context primarily enforced: the verification or falsification condition.

References

Cohen, Ariel & Krifka, Manfred. 2011. Superlative quantifiers as modifiers of meta-speech acts. In. Partee, B.H., Glanzberg, M., & Skilters, J. (Eds.) (2011). *Formal semantics and pragmatics. Discourse, context and models. The Baltic International Yearbook of Cognition, Logic and Communication,* Vol. 6 (2010). Manhattan, KS: New Prairie Press.

Cummins, Chris & Katsos, Napoleon. 2010. Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics*, 27: 271-305.

Geurts, Bart and Nouwen, Rick. 2007. "At least" et al.: The semantics of scalar modifiers. *Language*, 83: 533-559.

Geurts, Bart, Katsos, Napoleon, Cummins, Chris, Moons, Jonas, and Noordman, Leo. 2010. Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, 25: 130-148.

Koster-Moeller, Jorie, Varvoutis, Jason & Hackl, Martin. 2008. Verification procedures for modified numeral quantifiers, *Proceedings of the West Coast Conference on Formal Linguistics*, 27.

Krifka, Manfred. 1999. At least some determiners aren't determiners. In K. Turner (ed.), *The semantics/pragmatics interface from different points of view.* (= *Current Research in the Semantics/Pragmatics Interface* Vol. 1). Elsevier Science B.V., 257-291.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge, England: Cambridge University.

Musolino, Julien. 2004. The semantics and acquisition of number words: Integrating linguistic and developmental perspectives, *Cognition*, 93-1, 1-41.

Nouwen, Rick 2010. Two kinds of modified numerals. *Semantics and Pragmatics* 3 (3).

Zimmermann, Thomas Ede. 2000. Free choice disjunction and epistemic possibility, *Natural Language Semantics*, 8 (2000), 255 - 290.

Modelling Strategic Conversation: model, annotation design and corpus

Stergos Afantenos[1], Nicholas Asher[1], Farah Benamara[1], Anais Cadilhac, Cedric Dégremont[1], Pascal Denis[3], Markus Guhe[2], Simon Keizer[4], Alex Lascarides[2], Oliver Lemon[4], Philippe Muller[1], Soumya Paul[1], Vladimir Popescu[1],Verena Rieser[4] Laure Vieu[1]
1: IRIT, Univ. Toulouse, France
2: Department of Informatics, University of Edinburgh, U.K.
3: Mostrare, INRIA NLE, France
4: Herriott-Watt Univeristy, Edinburgh, U.K.

1 Introduction

A Gricean view of cognitive agents holds that agents are fully rational and adhere to the maxims of conversation that entail that speakers adopt shared intentions and fully aligned preferences– e.g. (Allen and Litman, 1987; Lochbaum, 1998). These assumptions are unwarranted in many conversational settings. In this paper we propose a different view and an annotation scheme for it.

We propose a game theoretic approach to conversation. While we assume like Grice that conversational agents are rational, agents talk to maximize their expected utility (a measure that combines belief and preference). Preferences together with beliefs guide conversational actions as much as they guide non linguistic actions. Conversations are dynamic and extensive games, and they have an in principle unbounded number of possible moves and no mandatory stopping points— you can, in some sense, always say anything, and you can always continue a conversation. The moves for each player consist in making a discourse contribution, which we finitely characterize using discourse structure in the sense of (Asher and Lascarides, 2003). Such discourse structures consist of discourse units linked to each via discourse relations like Elaboration, Question-Answer-Pair (QAP) and Explanation. In addition these discourse relations serve to link one participant's contribution to another; for instance, if one agent asks a question, another may respond with an answer, the two contributions then linked together by the relation QAP. Conversational participants are alternatively senders (S) or receivers of messages (R). S sends a signal s bearing in mind that receiver R has to figure out: (a) what is the message $m(s)$? What is S publicly committed to? (b) Is $m(s)$ credible or not? (c) Given a status for $m(s)$, what signal s' should R send in return? R now becomes sender and S, now the receiver, goes through the calculation steps (a)-(c). We assume that at least part of the conventional meaning of the signal is determined prior to game play. In calculation (a), R must calculate using a form of generalized signaling game what are the public commitments that S has made—these include not only the fixed semantics but also the implicatures that introduce discourse relations between contributions. Sometimes these involve strategic considerations: for instance, is S actually replying to the question asked in the prior turn or is she engaged in some other discourse move? If she is answering the question, is this something that S cannot plausibly later deny? (Asher and Quinley, 2011) argue that a trust game format is the right one for computing optimal moves in task (c).

(Traum and Allen, 1994) advocates a related view on which cooperativity is determined only by the social conventions guiding conversation, obligations that do not presuppose speakers adopt each other's goals (Traum et al., 2008). For us, the social conventions that are foundational on Traum's account are however themselves based on utility. Utility is also the basis for training agents to behave in a certain way through reinforcement learning for conversational agents (Frampton and Lemon, 2009).

2 Example negotiation dialogue

We provide a sample annotation of a negotiation dialogue (table (1) from our corpus, which consists of recorded chat negotiations taking place during on-line games of The Settlers of Catan,[1] a popular boardgame. The annotations are done

[1]See the original game on www.catan.com, adapted by us on homepages.inf.ed.ac.uk/mguhe/socl/

Speaker	Id	Turn	Dom. function	Rhet. function	Prefs
Euan	47	[And I alt tab back from the tutorial.]_1	OTHER		
		[What's up?]_2	OTHER	Result*(47_1,47_2)	
Joel	48	[do you want to trade?]	OFFER ⟨Joel,?,?,Euan⟩	Q-elab(47_2, 48)	
Card.	49	[joel fancies a bit of your clay]	STRAT.-COMMENT	Expl*(48, 49)	Pref(joel)
Joel	50	[yes]	OTHER	Ackn(49, 50)	
Joel	51	[!]	OTHER	Comment(50, 51)	
Euan	52	[Whatcha got?]	COUNTEROFFER ⟨Euan,?,?,Joel⟩	Q-elab([48-50], 52)	
Joel	53	[wheat]	HAS-RESOURCES ⟨Joel,wheat⟩	QAP(52, 53)	
Euan	54	[I can wheat for clay.]	COUNTEROFFER ⟨Euan,wheat,clay,Joel⟩	Elab([52,53], 54)	
Joel	55	[awesome]	ACCEPT(54)	Ackn(54, 55)	

Table 1: Example annotation, with offer arguments: offerer, requested resource, offered resource, receiver.

using the GLOZZ tool developed by the University of Caen.[2]

Our annotation model features both a discourse structure level (DS) and a dialogue act (DA) level, which categorizes elementary discourse units or EDUs, given by a pre-annotation, relative to their role in negotiations. DS encodes communicative functions of EDUs or clusters of EDUs using the relations of (Asher and Lascarides, 2003), similar to but more detailed than DAMSL's (Core and Allen, 1997). Unlike (Sidner, 1994), which also provides domain level acts for negotiations, our semantics for DAs does not assume Gricean cooperativity. Our DAs are: OFFER, COUNTEROFFER, STRATEGIC COMMENT a comment about a play in the game, OTHER. Each act also comes with an annotation of resources that are offered, requested, or simply possessed. With respect to the discourse relations, Expl* and Result* stand for "metalinguistic" relations: Result*(47_1, 47_2) means that the action described in 47_1 causes the speech act of asking the question in 47_2. Similarly, Expl*(48,49) indicates that Cardlinger explains why Joel asked the question in EDU 48. Q-elab is the relation of follow up question or Q-elab, and Ackn stands for the acknowledgment relation, while QAP stands for Question-Answer-Pair. The semantics for all these relations can be found in (Asher and Lascarides, 2003).

Our annotators received training over 22 negotiation dialogues with 560 turns. The inter annotation agreement at both EDU and rhetorical structure levels for this training will be used to refine the guidelines. In over 91 instances of doubly-annotated EDUs considered, we have a kappa of 0.54, a moderate level due to the very high number of "other" acts . For rhetorical structure, using an exact match criterion of success (easy to compute but harsher than necessary), we have a Kappa of 0.45. These figures are very preliminary.

References

J. Allen and D. Litman. 1987. A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11(2):163–200.

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

N. Asher and J. Quinley. 2011. Begging questions, their answers and basic cooperativity. In *LENLS 8*, Japan.

M. Core and J. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative actions in Humans and Machines*, pages 28–35, Cambridge, MA.

M. Frampton and O. Lemon. 2009. Recent research advances in reinforcement learning in spoken dialogue systems. *Knowledge Eng. Review*, 24(4):375–408.

K. E. Lochbaum. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.

C. Sidner. 1994. An artificial discourse language for collaborative negotiation. In *AAAI*, volume 1, pages 814–819. MIT Press, Cambridge.

D. Traum and J. Allen. 1994. Discourse obligations in dialogue processing. In *ACL94*, pages 1–8, Las Cruces, New Mexico.

D. Traum, W. Swartout, J. Gratch, and S. Marsella. 2008. A virtual human dialogue model for non-team interaction. In L. Dybkjr, W. Minker, and N. Ide, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*, pages 45–67. Springer Netherlands.

[2] www.glozz.org

Surprise, deception and fiction in children's Skype conferences

Thomas Bliesener
Dept. of Communication Science
University of Duisburg-Essen
thomas.bliesener@uni-due.de

Abstract

Synchronous multimedia teleconferences like Skype have limitations of reliability and quality, and they restrict basic communicative functions. One cannot monitor how one is heard and seen. Generally, the spiral of mutual awareness is impaired. Nevertheless, children create amazing scenes of apparent and hidden givings, of serious and fictitious, expectable and surprising. With dramaturgical talent, they compensate technological imperfection and bring about delightful and supportive encounters.

1 Impaired spiral of awareness in teleconferences

Communication requires means for awareness, anticipation and action. In many face-to-face situations, these conditions are sufficiently given and not reflected in detail. However, if communication is practiced over large distances by use of mediating computers and audiovisual technology, the required resources may be available only to a limited extent.

A typical limitation concerns the relation between the remote partner and the remote environment. If a local participant gets a video image of his remote partner, he may become aware of his partner‘s facial or bodily expressions. But he will not see what his partner is seeing and what he is doing in order to see it. Only additional video images of what his partner is looking at might remedy this shortcoming. In particular the participant is lacking an image of the remote display. So he cannot watch how his own video input is displayed. He has not enough cues to monitor how he is perceived. The same holds for auditive data. It is nearly impossible for the local participant to guess how he is being heard at the remote side. So he cannot reliably adapt his own voice to the conditions of his remote partner. Taken together, his awareness of his own efficacy is handicapped, and anticipations of his partner's reactions are impaired.

As a consequence, mutual awareness of awareness, which is given under the homogeneous conditions of face-to-face interaction, does not emerge in communication between separated places. In this case, the „spiral of awareness“ is precarious. This was already noticed in studies on CSCW. The concept of „meta-awareness“ (Fashchian, 2003) was to cover this phenomenon, similarly the concept of „translucence“ (Oemig, 2004; Szostek et al., 2008). However, technological *solutions* that allow to monitor how one‘s own input is being received at the remote end, are still missing. At least consumer equipment with Skype has no features to overcome these restrictions.

Instead, progress is made by users' competencies. Experienced users are aware of the shortcomings. They frequently perform cooperative sound and video checks, gather verbal feedback, and exchange screenshots of their video images (Bliesener, 2002).

Particularly interesting are sequences in teleconferences where action plans with higher order anticipations are pursued successfully. Examples were found in Skype conferences, recorded in a pilot project of the Dept. of Communication Science (Prof. Dr. H. W. Schmitz) [1]: In a paediatric ward for bone marrow transplantations, the children stay three months isolated in a sterile room. Very few visitors are admitted, and they must be cloaked in whole-body germ-free gear. The children's bridge to the familiar world outside are Skype conferences with audio, video, chat and games. The screen captures comprise 200 hours.

[1] „TKK-ELF: Telecommunication between Pediatric Cancer Patients and their Families, Teachers and Friends”. Report in (Bliesener et al., 2012)

2 Successful deceptions and surprises

Three examples of little orchestrations shall be described, illustrated and scrutinized with regard to their recipe for success.

a. A patient shows Yu-Gi-Oh cards to a peer. He approaches them to the camera so that they look larger and, as a side effect, hide his face. The first and the second presentation appear like rehearsals for targeting the card into the shot. The third time, he retracts the card exactly on the axis between the camera and his hidden face. Suddenly he takes it aside, thereby unveils his face like starting a Punch and Judy show, smiles demonstratively and makes a sound like „Hello“. His formula for success is the use of repetition for both the creation of an expectation and the rehearsal of an action to be performed “blindly”.

b. A patient announces “Now I am going to smash the cam”. He moves back one arm, forms a fist and pushes it towards the camera. There, he covers the lens with his palm and, protected by invisibility, turns off the camera. Simultaneously he makes a sound like broken glass. His formula for success is creating the expectation of a result of his action, then using different actions to create the semblance of the expected result. This is the same principle as in a piece of legerdemain.

c. A friend of a patient changes his audio configuration so that he can listen by boxes instead by headset. He puts his headset beyond the camera angle and ensures by a sound check that the associated microphone still captures his voice. Invisibly to the patient, he lays his cat on the headset. When the cat starts purring, the patient is highly surprised and delighted. A camera shift to the cat with headset unravels the mystery. Here, the recipe for success is a combination of precise knowledge of the audio features and a purposeful reduction of an object to audio only. It is technological sophistication paired with dramaturgical minimalism.

3 Résumé: Factors of telecommunicative competence

Success in creating surprise and delight via Skype comes about by at least three factors:
- *sophisticated knowledge* of technological features within the limits of their disposability,
- *trained action* sequences for „blind“ performance with insufficient or no sensomotoric feedback,
- *artful creation*, condensation and implementation of peculiar references to shared knowledge.

Taken together, these factors resemble much what a dramaturgical talent is expected to comprise. This finding should not be a big surprise. From the social viewpoint of a user, technology for synchronous multimedia telecommunication offers just a new stage with new requisites. Talented players will make the most of it, since this is the core of their art: showing *a world in a grain of sand*.

4 Conclusion: Competence as a compensation for technological imperfection

In ratings of the Skype recordings by neutral observers, such examples of interactions with deception, surprise, pantomime etc. are considered as highlights of interpersonal closeness. The children’s unfolding of dramaturgical competencies seems to work as a compensation for the technological restrictions. It does do so simply by their joy and art of playing,. If children’s inclination is not channelled to pre-defined and pre-fabricated digital games with fixed rules and scores, it can have the power to turn sober and imperfect teleconferences into playful interactions with a considerable potential of social support. As a consequence, future training and support for Skype should encourage adults, too, to reawaken the phantasy and creativity of their childhood, to *play* with features and to *parody* failures. They might learn from their children how much fiction can improve experience.

References

Agnieszka M. Szostek et al. 2008. Understanding the Implications of Social Translucence for Systems Supporting Communication at Work. CSCW’08, November 8-12, San Diego, 649-658.

Babak A. Fashchian. 2003. Presence Technologies for Informal Collaboration. In: Riva, G. et al (eds.): Being There. Trondheim, 209-222.

Christoph Oemig. 2004. Awareness Management in Socially Translucent Systems, Furtwangen, 34-35.

Thomas Bliesener. 2002. Taste the difference. Training and optimization of synchronous, audiovisual, cooperative telelearning. In: Proceedings of E-Learn 2002, Montreal, Canada, Oct. 15-19, 2002. CD-rom:\procbook.pdf, 1199-1201.

Thomas Bliesener et al. 2012. Telekommunikation gegen Isolation. Aachen (forthcoming)

A Multi-threading Extension to State-based Dialogue Management

Tina Klüwer & Hans Uszkoreit
German Research Center for Artificial Intelligence
Alt-Moabit 91c, 10559 Berlin, Germany
tina.kluewer@dfki.de, hans.uszkoreit@dfki.de

1 Introduction

This paper describes a dialogue manager, which provides support for multiple, interleaved conversation threads. Multi-threaded dialogues are frequently initialized by humans (Shyrokov et al., 2007), (Yang et al., 2008). Interleaved dialogue threads differ from embedded dialogue threads insofar that they allow for threads being alternated entangled. Although multi-threaded conversations are a frequent human behavior, support for multi-threaded conversations in dialogue systems is very rare. One example is (Lemon et al., 2002), who describe a possibility to integrate multi-threading into an Information State Update model. However, (Yang et al., 2008) criticize (Lemon et al., 2002), because they neglect to signal conversation switches made by the system. The system described in (Nakano et al., 2008) is able to manage multiple tasks through several expert components for every task. However, experts cover fine-grained tasks such as "understanding a request for weather information". They do not capsulate substructures of a dialogue and are therefore not comparable to conversation threads.

This paper presents a state-based dialogue manager, which supports multi-threaded behavior and offers conversation switch markers.

2 Multi-threaded Dialogue Support

Input to the dialogue system is by default interpreted in the context of the currently active dialogue thread. If the current thread is found to be inappropriate, the dialogue manager needs to replace the active thread with either an until now inactive thread or with an active but paused thread. Dialogue Management is based on a finite-state graph. The finite-state automaton is described by a hierarchical state-transition diagram including Harel's state charts. In our dialogue manager conversation threads are special types of supernodes.

Conversation threads can occur in three different conditions: active, paused and inactive. Analog to the activation, termination and pausing of conversation threads, the underlying graph interpreter activates, terminates and pauses the belonging thread supernodes.

Empirical research has stated that the change of conversation threads by the system can easily become confusing to the user (Heeman et al., 2005), especially if the system does not provide a discourse marker to notify the change.

Therefore, the described system provides "bridging utterances" to indicate a thread switch. They consist of two parts: The first one is a more general reference to the newly activated or reactivated thread (mostly through verbalizing the topic of the selected thread), the second one the repetition or rephrasing of the last utterance which was made by the system to reestablish common ground.

3 Selection of Dialogue Threads

In the **system initiative** scenario the dialogue manager has to decide which dialogue thread constitutes an appropriate continuation of the conversation, e.g. after a dialogue thread was finished and the conversation pauses. The system can choose between reactivating a paused dialogue thread or activating a new thread. Two criteria are used for the selection: time information (since when an active thread is paused) and importance information (how important is the thread for the overall conversation).

For **user initiative**, all incoming user utterances are by default at first processed in the context of the currently active thread. If the active thread fails in offering a valid transition to a new state, the dialogue manager selects a dialogue thread which fits to the incoming utterance. This can be a paused thread or a thread which has to be freshly initialized. The selection process is led by the matching values for the topic of the utterance as well as the recognized dialogue act and domain. If more than one thread with matching values for topic, domain and a valid dialogue act can be found, the selection process continues with the measures for importance and time.

4 Evaluation of Thread Selection

The dialogue manager was evaluated through conversation logs from user experiments. Each utterance in the conversation logs was manually annotated with thread function information. Thread functions include the opening of new threads, the reinitialization of paused threads and the selection of threads according to user utterances.

Unfortunately, in the evaluation experiments the users did not make use of the interleaved dialogue possibilities, but only used embedded dialogue threads. However, since the system does not differentiate between embedded or interleaved threads, we expect the system to also provide good support for interleaved conversations.

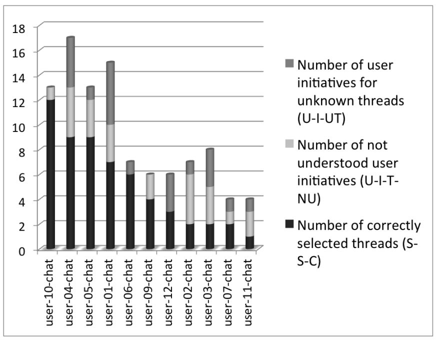


Figure 1: User initialized dialogue threads and system reaction

In general, the evaluation shows that the thread selection works very well.

Figure 1 shows the division of the system's thread selections as reaction to thread initialization by the users. There were no incorrectly selected threads by the dialogue manager, but a number of problems originating from failures of the NLU component of the system (U-I-T-NU) and some cases in which the users initialized dialogue threads unknown to the system (U-I-UT). In total 23 of 102 user's attempts to initialize new dialogue threads were not understood by the input analysis (25,48%).

Thread selection for system initiative also works very well. There were only 16 errors in 157 thread selections. Most of the errors (13 of 16) are caused by a missing behavior in the selection algorithm, which did not consider the number of already uttered rejections by the user.

The system reinitialized 63 paused threads either because of a user utterance or as system initiative. The number of reinitialized threads per conversation differs from 29 (the highest number) to 8 (the lowest number of reinitialized threads). All threads for reactivation were correctly selected.

Acknowledgments This work was supported by the project "Sprinter", funded by the German Federal Ministry of Education and Research (BMBF) under grant number 01IS12006A.

References

Peter A. Heeman, Fan Yang, Andrew L. Kun, and Alexander Shyrokov. 2005. Conventions in human-human multi-threaded dialogues: a preliminary study. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 293–295, New York, NY, USA. ACM.

Oliver Lemon, Alexander Gruenstein, Alexis Battle, and Stanley Peters. 2002. Multi-tasking and collaborative activities in dialogue systems. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue - Volume 2*, pages 113–124, Stroudsburg, PA, USA. ACL.

Mikio Nakano, Kotaro Funakoshi, Yuji Hasegawa, and Hiroshi Tsujino. 2008. A framework for building conversational agents based on a multi-expert model. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. ACL.

Alexander Shyrokov, Andrew Kun, and Peter Heeman. 2007. Experimental modeling of human-human multi-threaded dialogues in the presence of a manual-visual task. In *Proceedings of the SIGdial 2007*, Antwerp, Belgium.

Fan Yang, Peter A. Heeman, and Andrew Kun. 2008. Switching to real-time tasks in multi-tasking dialogue. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pages 1025–1032, Stroudsburg, PA, USA. ACL.

Negotiation for Concern Alignment in Health Counseling Dialogues

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

Yosuke Matsusaka
Advanced Institute of Science and Techonology, Japan
yosuke.matsusaka@aist.go.jp

Abstract

'Concern Alignment in Conversations' project aims to elucidate interplay between rational agreement seeking and affective trust fostering in conversation. The project is based on empirical analyses of real life conversation data in medical counseling domain, and development of computational models.

1 Introduction

Conversational interactions in real life, in many cases, are purported to form a consensus on something among conversational participants. Consensus can be conceived as a formation of shared commitments on certain choice of future joint actions by a group of people. However, process is as important as outcome in consensus-building. Consensus decision making process should be collaborative and participatory so that 'consensus seeks to improve solidarity in the long run' and participants 'shape it into a decision that meets the concerns of all group members as much as possible'(wikipedia,). This affective process, which eventually leads to fostering of trust, is in parallel with the rational process of seeking agreement. 'Concern Alignment in Conversations' project aims to elucidate this interplay between rational agreement seeking and affective trust fostering processes in conversation, based on analyses of real life conversation data.

2 Research issues

Descriptive concepts: Identify the descriptive concepts needed to adequately capture the processes in which conversation participants work together to promote 'affective solidarity.'

Trust-related dialogue acts: Identify dialogue acts performed by conversational participants that contribute to 'affective solidarity.'

Agreement/Trust interrelationship: Identify relationships between rational agreement seeking and affective trust fostering.

Mapping: Devise a mapping between surface linguistic acts and trust-related dialogue acts.

3 Concern alignment model

We picture a consensus decision-making process as consisting, conceptually, of two parts (Fig.1)(Katagiri et al., 2011; Katagiri et al., 2012). First, once a group of people are placed in a situation in which they need to make an agreement on their joint course of actions (*issues*), they start by expressing their value judgments on properties and criteria on actions each considers significant (*concerns*). After they share their concerns, they start proposing and negotiating on concrete choice of actions (*proposals*). When we decide on which restaurant to go for lunch, we first discuss on what restaurant properties we put most priorities, e.g., price, location, cuisine etc. We, then, start talking about actual restaurants.

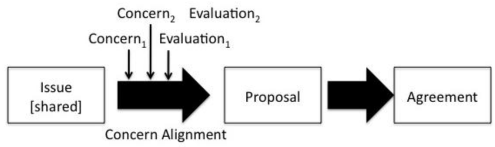


Figure 1: A schematic diagram of the concern alignment process

4 Dialogue data & analysis

Data

We chose dialogues in medical counseling sessions where people diagnosed as obese (metabolic syndrome) see expert nurses to get advises on their daily life management. Screening and counseling for obese people have been mandated by the Japanese government, and counseling services are now regularly offered by many hospitals. The patients are often reluctant to follow nurses' advices, and it is important for nurses to establish rapport with patients to enforce their advices. We have collected a total of 9 sessions, about 5 hours of dialogues on video. All the sessions were transcribed.

Analysis

We have observed several dialogue acts for concern alignment.

Concern introduction: Since an issue has mostly been shared in our setting, e.g., deciding on plans to counter the patient obesity, nurses and patients express their concerns in the form of broad category of actions to take, e.g, whether to reduce calorie intake, to increase exercise, to stop smoking or to rely on other methods.

Evaluative responses to concerns: Once a concern is introduced, the interlocutor expresses a positive or negative evaluations of it. The expression can be either linguistic or non-linguistic. Negative expressions often take the form of expressing conflicting concerns.

Incremental alignment: When a concern is positively evaluated by the interlocutors, they can now proceed to discuss more concrete proposals based on the shared concern. When a concern is negatively evaluated, they have to negotiate. The negotiation often takes the form of focusing, e.g., introducing a related but slightly modified concern. Figure 2 is an abbreviated excerpt showing a focusing type incremental alignment, in which initial concern of 'eating less' was rejected by a counter introduction of 'bicycle,' which is modified by 'commute,' which leads to a proposal of 'bicycle commute.'

A	<u>Body weight reduction</u>, reduction of fat, is effective in many cases. So, you should consume about 230 kCal a day, maybe, by <u>eating a little bit less</u> than now, I think.
B	Uh huh.
	...
B	I hear people talk about walking, but how about <u>bicycle</u>?
A	Bicycle is of course fine. ··· Do you ride a bicycle?
B	Well, not much exercise. I commute by car.
A	Uh huh.
B	I wonder how long it takes on foot to my office.
	...
	It takes more than 30min., so I think it's a bit too much for commute.
A	Yeah, Well, it'll getting cooler
B	Yeah
A	in the morning and at night.
B	It would be nice if I can exercise by bicycle, may be.
A	Yes. Do you get back home early or late?
B	It's late.
	...
A	Then, if it's OK for you,
B	uh huh
A	how about <u>bicycle</u>
B	Bicycle
A	<u>commute</u>?
B	I think <u>I will try bicycle commute</u>.
A	Yeah, that will be good.

Figure 2: Concern alignment by focusing.

References

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 Emonoto, Yasuharu, and Yosuke Matsusaka. 2012. An analysis of mutual trust formation in health guidance dialogues. Technical Report SIG-SLUD-B103: 89-94, Technical Report of Japanese Society of Artificial Intelligence.

wikipedia. https://secure.wikimedia.org/wikipedia/en/wiki/Consensus_decision-making. accessed on 10 June, 2012.

Exhuming the procedural common ground: partner-specific effects

Gregory Mills
University of Edinburgh
gmills@staffmail.ed.ac.uk

Introduction

One of the central findings in dialogue research is that interlocutors rapidly converge in their use of referring expressions, and that this convergence is driven by the interaction: if interlocutors are able to provide each other with communicative, turn-by-turn feedback, this leads to the quicker development of representations that are more concise (Clark, 1996), more compositional (Garrod et al, 2007), more systematic and more abstract (Healey, 1997), and are also more tailored to specific conversational partners (Healey and Mills, 2006; Brown-Schmidt et al 2007).

1 Procedural co-ordination in dialogue: Complementary contributions

However, in addition to co-ordinating on the content of referring expressions, interaction in dialogue also requires procedural co-ordination: interlocutors must co-ordinate on the sequential and temporal unfolding of their contributions. Empirical studies of conversational interaction have demonstrated that procedural co-ordination is underpinned by interlocutors' use, not of the same, but of different kinds of contribution. For example, questions are usually followed with answers, not with another question, requests are usually followed with compliance, not with counter-requests, praise is usually followed with self-denigration, and offers with acceptance. These adjacency-pairs (Schegloff, 1992) are conventions which operate normatively, and consist of a first-pair part and a second-pair part, performed by different speakers. A central feature is that their successful use typically requires interlocutors to perform different and complementary contributions on subsequent turns. However, both conversation analytic and cognitive studies of interaction have treated these adjacency pairs as already shared and known to be shared by interlocutors, and do not study how interlocutors converge on them in the first place. It is also unclear whether convergence is driven primarily by egocentric processes (i.e. relatively low-level routinization), or whether interlocutors readily associate these conventions with specific conversational partners.

2 Alphabetical sorting task

To address these questions, we report a collaborative 3-participant task which presents participants with recurrent procedural co-ordination problems. Participants communicate via a text-based chat tool (Healey and Mills, 2006). Each participant's computer also displays a task window containing randomly generated words. Solving the task requires participants to combine their lists of words into a single alphabetically ordered list. To select a word, participants type the word preceded with "/". To ensure collaboration, participants can only select words displayed on the other participant's screen and vice versa. Note that this task is trivial for an individual participant. However, for groups of participants, this task presents the co-ordination problem of interleaving their selections correctly: participants cannot select each other's words, words can't be selected twice, and words need to be selected in the correct order (See Mills, 2011 for a similar task).

To examine whether participants readily associate these conventions with specific conversational partners, the 3 participants were divided into a main dyad and a second side-participant. The task was configured such that at key moments in the development of the conventions, the side-participant is only required to observe the interaction, but does not directly participate in establishing the conventions.

To test for partner-specific effects, we drew on the method of (Healey and Mills, 2006) of using a chat server to intercept and selectively manipulate participants' turns in real-time. This technique is used to generate artificial clarification requests that query the procedural function of participants' turns. The apparent origin of these clarification requests is manipulated to appear as if they originate from either of the 2 other participants (Main Dyad vs. Side participant).

Comparison of the responses to these two types of artificial clarification request allows direct testing of the hypothesis that interloctors associate the co-ordination they achieve with specific conversational partners.

3 Results

We demonstrate that participants' responses to these clarification requests provide strong evidence of interlocutors associating procedural conventions with specific partners. Despite the clarification requests having exactly the same surface form (all that differs is their apparent origin), responses to both types of clarification are treated differently: Participants are slower to respond to clarification requests from the side-participants, their responses are also longer, contain more self-corrections, and they also subsequently make more mistakes in the task. Drawing on global interaction patterns in the task, we also demonstrate that these partner-specific effects are sensitive to the specific sequential location in the dialogue where problematic understanding is signaled.

4 Complementarity, Convergence and Conventionalization.

We argue that focusing on procedural co-ordination suggests a more nuanced view of convergence in dialogue. The rapid development of conventions consisting of complementary contributions suggests that the global development of procedural co-ordination that occurs over the course of the interaction involves systematic *divergence* at a local turn-by-turn level. Drawing on participants' patterns of interaction in the task, we argue that this differentiation is indicative of a greater "forward momentum" in the interaction, as it indicates that participants have converged on what the next relevant step is in the dialogue. By contrast, high levels of local convergence between turns is indicative of lower levels of communicative success, as this typically indicates that interlocutors have halted the interaction in order to identify and resolve problematic understanding.

We also argue that the finding of partner-specific effects also points towards differentiation and divergence occurring at more global levels of interaction – although all the participants are exposed to exactly the same communicative behaviour from each other (they all see the same interaction unfold on the screen), as they become more co-ordinated in the interaction, the main dyads and the side-participants systematically adopt different procedural conventions that become progressively complementary as their roles diverge.

5 References

Brown-Schmidt, S. & Tanenhaus, M.K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. Cognitive Science, 32, 643-684.

Clark, (1996). Using Language. CUP. Cambridge

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of Representation: Where Might Graphical Symbol Systems Come From? *Cognitive Science:* 31(6), 961–987.

Healey, P.G.T. (1997). “Expertise or expert-ese: The emergence of task-oriented sub-languages” Proceedings of the 19th Annual CogSci Meeting. Stanford University, CA.

Healey, P. G. T. & Mills, G. (2006). Participation, precedence and co-ordination. In Proceedings of the 28th Conference of the Cognitive Science Society, Canada.

Mills, G. J..(2011). “The emergence of procedural conventions in dialogue” Proceedings of the 33rd Annual CogSci Meeting. Boston, USA.

Pickering, M. J. and Garrod, S. (2004). Towards a Mechanistic psychology of dialogue. *Behavioural and Brain Sciences*27(2):169–190.

Schegloff, E. A. (1992) Repair after next turn *AJS* 97(5)

Opponent Modelling for Optimising Strategic Dialogue

Verena Rieser, Oliver Lemon, and Simon Keizer
School of Mathematical and Computer Sciences (MACS)
Heriot-Watt University
Edinburgh EH14 4AS, UK.
{v.t.rieser, o.lemon, s.keizer}@hw.ac.uk

Abstract

Within the wider context of the STAC project, we are developing new models of non-cooperative strategic conversation. We concentrate on learning optimised negotiation strategies (such as deception and information hiding) from real data collected in the domain of "Settlers of Catan", a multi-player board game. This paper illustrates how multi-agent reinforcement learning techniques can be used to model strategic dialogue behaviour. In particular, we discuss novel probabilistic models, called "interactive POMDPs", which combine game theoretic opponent modelling with Partially Observable Markov Decision Processes.

1 Introduction

Within the wider context of the STAC project (2012-2017) we are developing models of non-cooperative strategic conversation[1]. While other partners explore the linguistic and game-theoretic underpinnings of non-Gricean behaviour (Asher and Lascarides, 2008), we focus on learning negotiative dialogue strategies from real data.

The STAC project is collecting data on human trading strategies while playing a modified online version of the board game "Settlers of Catan" (Thomas and Hammond, 2002) where players negotiate trades via a chat interface (Guhe and Lascarides, 2012).

In the following we illustrate how multi-agent reinforcement learning (RL) can be used to optimise strategic trading actions such as deception and information hiding. Previous work has explored single-agent RL for negotiation strategies (Georgila and Traum, 2011; Heeman, 2009), using very limited amounts of data and limited strategic reasoning.

2 Opponent Modelling for Strategic Trading

Single-agent RL approaches were successfully applied to handle uncertainty in Spoken Dialogue Systems, see e.g. (Rieser and Lemon, 2011). However, when considering non-cooperative bargaining domains such as resource negotiation in Settlers, a new type of uncertainty has to be modelled: agents can lie, deceive, bluff, and hide information (Osborne and Rubinstein, 1990). This type of partial observability falls outside the scope of current Partially Observable Markov Decision Processes (POMDPs) approaches to dialogue (Williams and Young, 2007), which focus on uncertainty derived from speech recognition errors.

Examples from an initial data collection (Guhe and Lascarides, 2012) show that human Settlers players employ elaborate strategic conversational moves: On the one hand, players deflect by providing misleading implicatures (Example 1b), hold back information by not answering a question (1c), or tell explicit lies. On the other hand, seemingly cooperative strategies, such as volunteering information, can be observed (Guhe and Lascarides, 2012). Furthermore, offers as in Example (1a) are also often under-specified or "partial", i.e. instead of explicitly specifying how many resources are offered and how many are needed, this information is only revealed strategically in the course of the dialogue.

(1)
a. A: Do you have rock?
b. B: I've got lots of wheat [in fact, B has a rock]
c. C: [*silence*]

[1] http://www.irit.fr/STAC/

In order to account for this type of strategic dialogue behaviour, we are exploring novel probabilistic models which combine game-theoretic and POMDP control strategies. In game theory, the process of inferring strategies of other players is also known as "k-level thinking" or opponent modelling (Leyton-Brown and Shoham, 2008). The RL community has adapted these ideas for multi-agent adversarial learning using minimax Q-learning (Littman, 1994) or interactive Partially Observable Markov Decision Processes (iPOMDPs) (Gmytrasiewicz and Doshi, 2005). We extend iPOMDPS for extensive-form games with sequential actions, see (2).

$$I\!-\!POMDP = < IS_i, A_i, T_i, \Omega_i, O_i, R_i > \quad (2)$$

$IS_i = S \times M_j$ is a set of interactive states, where S is the set of states of the physical environment, and $\{M_{j...m}\}$ is the set of possible models of agents $j...m$. $\{A_i\}$ describes agent i's set of actions. $T_i : IS \times A_i \times IS \rightarrow [0, 1]$ is a transition function which describes results of an action. Ω_i is the set of observations the agent i can make. $O_i : IS \times A_i \times \Omega_i \rightarrow [0, 1]$ is the agent's observation function which specifies probabilities of observations given agents' actions and resulting states. Finally, $R_i : S \times A_i \rightarrow R$ is the reward function representing agent i's preferences.

By formulating an "interactive state" which includes explicit possible behavioural models of other agents, iPOMDPs recognise that agents are not playing against distributions like in single-agent RL, "but other players who understood the rules and were prepared to leverage them against slower players" (Wunder et al., 2011).

3 Action Set for Learning

We will employ iPOMDPs to model and reason about hidden states and strategic conversational behaviour of other players. In particular, we aim to learn optimal behaviour for the following decisions:

1. How to make a strategic offer and how much information do I expose?
2. How do I reply to an offer and how sincere is my reply?

For learning the first decision, we modify the JSettlers system, where an artificial trading agent can only pose a fully specified offer ("*[A], I'll give you 1 wheat for 1 sheep.*") via a graphical interface. We will handle partial offers (see Section 2) as well as disjunctive offers and requests (e.g. "wheat or sheep").

For optimising the latter decision, STAC has developed an annotation scheme which distinguishes between observable replies and their sincerity based on the logged game state. We plan to evaluate the success of these strategic dialogue capabilities against the original graphical version.

Acknowledgments

The research leading to this work has received funding from the European Community's FP7 programme (FP7/2011-16) under grant agreement no. 269427 (STAC). `http://www.irit.fr/STAC/`

References

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

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

Gmytrasiewicz, Piotr J. and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79.

Guhe, Markus and Alex Lascarides. 2012. Trading in a multiplayer board game: Towards an analysis of non-cooperative dialogue. In *Proc. of CogSci*.

Heeman, Peter. 2009. Representing the reinforcement learning state in a negotiation dialogue. In *Proc. of ASRU*.

Leyton-Brown, Kevin and Yoav Shoham. 2008. *Essentials of Game Theory: A Concise, Multidisciplinary Introduction*. Morgan & Claypool Publishers.

Littman, Michael L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proc. ICML*, pages 157–163.

Osborne, Martin J. and Ariel Rubinstein. 1990. *Bargaining and markets*. Academic Press.

Rieser, Verena and Oliver Lemon. 2011. *Reinforcement Learning for Adaptive Dialogue Systems*. Theory and Applications of Natural Language Processing. Springer.

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

Williams, J. and S. Young. 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language*, 21(2):231–422.

Wunder, Michael, Michael Kaisers, Michael Littman, and John Robert Yaros. 2011. Using iterated reasoning to predict opponent strategies. In *The 10th International Conference on Autonomous Agents and Multiagent Systems*.

What Should I Do Now? Supporting Progress in a Serious Game.

Lina M. Rojas-Barahona
INRIA/LORIA Nancy
lina.rojas@loria.fr

Claire Gardent
CNRS/LORIA Nancy
claire.gardent@loria.fr

Abstract

We present two dialogue systems developed to support chating with French speaking virtual characters in the context of a serious game: one hybrid statistical/symbolic and one purely statistical. The player is guided in the quest by different interactions including twelve distinct dialogs with different virtual characters.

1 Introduction

We present two state-of-the art dialogue systems. The first system (H) is a hybrid approach that combines an information-state dialogue manager (Burke et al., 2003) with a classifier for interpreting the players' phrases. The second system (QA) is a question/answering model which predicts the system dialog move based on a player's utterance (Gandhe et al., 2011). Both systems use a generation-by-selection strategy [1]. While previous work has focused on relatively short dialogs in a static setting, we consider longer interactions where dialogs occur in a setting that dynamically evolves as the game unfolds. Our conversational agents interact in French with virtual characters (VCs) in the context of the serious game Mission Plastechnologie (MP). In Section 2, we present the MP game and the dialogue strategies. Section 3 presents the two dialog systems. Finally we present a brief discussion in Section 4.

[1]The system's utterances are selected from a corpus annotated with system dialog moves

2 Dialogues in the MP Game

The MP game is a multi-player quest where 3 teenagers seek to build a joystick in order to free their uncle trapped in a video game [2]. To build this joystick, the player (who alternatively represents anyone of these three teenagers) must explore a plastic factory and interact with different VCs through twelve distinct dialogs, each of them occurring in a different part of the virtual world with different goals to be achieved. We identified four main dialog strategies, one general and three specific, and used these to define the plans guiding the rule-based engine in the H system. The general strategy is for information-seeking, the player looks for information about how to achieved some game-goals and the system provides this information. The specific strategies covers a request for pursuing a goal, a confirmation that a task has been accomplished and a negociation step.

3 Dialogue Systems

The game and the two dialog systems built were integrated as agents within the Open Agent Architecture (Cheyer and Martin, 2001). Both systems access a database for starting the appropriate subdialogs at the appropriate place in the virtual world; and for simultaneously storing all interactions in the database.

3.1 The Hybrid Dialogue System

The H system combines an interpreter, a rule based, Information State Approach dialog manager,

[2]The MP game was created by Artefacto, `http://www.artefacto.fr/index_ok.htm`

a generator and the game/dialog communication components.

The Interpreter Module The interpreter is a Logistic-Regression classifier which maps the player's utterance to a dialog move. The features used for training are the four previous system moves and the words filtered by tf*idf (Rojas-Barahona et al., 2012a). The best prediction given by the classifier is matched against the expected move determined by the *last move* stored in the information-state. In case of a mismatch, the interpreter selects a valid move in the current context and updates the information state with this move.

The Dialog Manager We designed a plan for each dialog strategy and extended Midiki (Burke et al., 2003) to support the OAA architecture and access the database with the configuration of the different dialogs in the game. Each time a new dialog is launched, the information state is loaded with the corresponding context (e.g., speakers, list of goals to be discussed) and the plan modeling the corresponding dialog strategy. We implemented a set of update and selection rules for integrating players' moves, handling the information-state and for preparing the agenda according to the plan. Once the system move has been selected, the Generator searches an appropriate verbalisation.

The Generator Given the system dialog move predicted by the dialog manager and the identifier of the current dialog, the generator picks randomly from the annotated corpus an utterance with these dialog move for that dialog identifier. In addition, propositional questions (i.e., proposals by the system to discuss additional topics) were annotated with their respective dialog goals. For example, Samir's sentence: *Are you interested in hearing about my job, the people that work here or the security policies?*, was annotated with the goals: *job*, *staff* and *security_policies*. For these dialog acts, the generator checks the list of current missing goals so as to retrieve an appropriate propositional question. In this way, the system can coherently direct the player by suggesting possible topics without using vague and repetitive sentences such as *Would you like to know more?*

3.2 The QA System

The QA system combines a Logistic-Regression classifier that matches players' turns to system dialog moves with the same generation-by-selection algorithm used in the H system. This classifier has been trained with the same features used for training the interpreter in the H system. Like the H system, the QA dialog system maintains a constant interaction with the game to allow for the game coherence and continuity.

4 Discussion

We have presented two system architectures for conversational agents situated in a serious game. While the QA system simply matches the player's input to a system response, the H system has a much more elaborate dialog managment policy including re-interpretation and the use of game and dialog history information, to guide the dialog. As a result, the QA dialogs are generally more spontaneous, giving the player more initiative whereas the H system enforces a more System-Driven dialog strategy thereby guiding the player through the game. A detailed comparison and evaluation of these two systems has been reported in (Rojas Barahona et al., 2012b)[3]

References

C. Burke, C. Doran, A. Gertner, A. Gregorowicz, L. Harper, J. Korb, and D. Loehr. 2003. Dialogue complexity with portability? In *Proceedings of the HLT-NAACL 2003 workshop on Research directions in dialogue processing.*

A. Cheyer and D. Martin. 2001. The open agent architecture. *Autonomous Agents and Multi-Agent Systems*, 4(1-2):143–148.

S. Gandhe, A. Taylor, J. Gerten, and D. R. Traum. 2011. Rapid development of advanced question-answering characters by non-experts. In *Proceedings of SIGDIAL*, pages 347–349, Portland, Oregon, USA.

L. M. Rojas-Barahona, A. Lorenzo, and C. Gardent. 2012a. Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents. In *Proceedings of LREC*.

L. M. Rojas Barahona, A. Lorenzo, and C. Gardent. 2012b. An end-to-end evaluation of two situated dialog systems. In *Proceedings of SIGDIAL*, pages 10–19, Seoul, South Korea.

[3]The research presented in this paper was partially supported by the Eurostar EmoSpeech project.

“The hand is not a banana”
On Developing a Robot’s Grounding Facilities

Julia Peltason, Hannes Rieser, Sven Wachsmuth und Britta Wrede
Bielefeld University

We investigate human-robot interaction and human-robot communication (HRI, HRC) with a robot named Flobi trained to acquire category terms like “banana” or “pine-apple”. More precisely, we present a system in which a robot and a human WOZ observer interact on this purpose. For ease of reference we call the whole system robot.
Intuitively, we consider a sort of tutorial dialogue between a human user and the robot. We do a deep evaluation of such a dialogue relying on standard paradigms, CA, dialogue theory, pragmatics, and formal theories of public information. Consequently, the HRC is seen as a datum in the ethno-methodological sense. As standards of comparison we use several notions set up or extended by H. Clark (Clark 1996, Clark and Marshall 1981) such as “action ladder”, common ground (CG), grounding (G), and mutual information. These notions all derive from human-human interaction (HHI), hence we call them STandard CG and STandard G. We contrast STandard CG and STandard G with Foundational CG and Foundational G needed in our tutorial dialogues. There is a difference between these two concepts, standard and foundational: Foundational CG and Foundational G deal with rooting a concept. They are hence at the basis of establishing a convention of use. To grasp that in David Lewisian terms one can say that Foundational CG and Foundational G mark the very beginning of a convention for the use of a category term, where a dyad, a robot and a human, starts to form a group initiating a convention of use. In other words, we deal with the building up of a precedent, a case not treated in Lewis’ “Convention” (1969).

We show in our talk how Foundational CG and Foundational G can fail or be achieved in the human-robot interaction. For this purpose we present two case studies, “The hand is not a banana” (fail) and “The pine-apple” (success). Both studies relying on standard paradigms and Clark’s “action ladder” as reference points show general problems arising in HRC. We point out which of these are due to the set up of the system, for example to the ASR and the grammar used or to the behaviour of the human user interacting with the robot. On the user’s side problems might arise due to his style of pointing (deixis) or to his use of politeness conventions not accounted for in the set-up of the current system (see however Peltason et al., 2012). From the case studies we can derive a preference list of mechanisms to be generally observed in the construction of robots, a FCG and FG scale. Most communication problems

arising in human-robot dialogues are well known from semantics, pragmatics and dialogue theory. Others do simply not arise in natural data since they are due to the workings of the system, for example, to its parsing component or its interaction patterns.

In contrast, from investigation of HRC we also get information we cannot normally access in HH communication. A case in point is the possibility to inspect the robot's mind, since we have access to the results of the ASR decodings and the content of deictic acts and can trace which internal state led to which verbal behaviour. This constellation has an interesting methodological side-effect: Especially the success case in the "pine-apple" study shows that in order to derive a working proposition one need not always rely on a standard "action-ladder" and that a proposition can be derived using multi-modal information.

Bibliography

Barwise, J. (1989). On the Model Theory of Common Knowledge. In :The Situation in Logic. CSLI Lecture Notes. 201-221.

Clark, H. (1996). Using Language. CUP, Cambridge.

Clark, H. and Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick et al. (Eds.), *Perspectives on socially shared cognition*. Washington, DC: APA Books, 127-149.

Clark, H. and Marshall, C. (1981). Definite reference and mutual knowledge. In *Elements of Discourse Understanding* (ed. A. Joshi et al.). Cambridge University Press.

Garrod, S. and Pickering, M. (2004). Toward a mechanistic psychology of dialogue. BBS 27, 169–190.

Ginzburg, J. (2012). The Interactive Stance. Meaning for Conversation. Oxford: Clarendon Press.

Lewis, D. (1969). *Convention*. Cambridge, Mass.

Peltason, J., N. Riether, B. Wrede, and I. Lütkebohle (2012). Talking with Robots about Objects: A system-level evaluation in HRI", *Proc. ACM/IEEE Conf. on Human-Robot-Interaction*, Boston 2012.

Traum, D. (1999). Computational models of grounding in collaborative systems. In *Working notes of AAAI Fall Symposium on Psychological Models of Communication*, pp. 124–131.

Quantitative experiments on prosodic and discourse units in the Corpus of Interactional Data

Klim Peshkov, Laurent Prévot, Roxane Bertrand, Stéphane Rauzy, Philippe Blache
Laboratoire Parole et Langage
Aix Marseille Université & CNRS
Aix-En-Provence, France
`firstname.lastname@lpl-aix.fr`

1 Introduction

The recent years have seen growing the number initiatives related to the interface between syntax, prosody and discourse. While in English the computational counterpart of this perspective has been largely advanced both from more formal modeling and machine learning perspectives, in French the situation is much less clear. Some automatic tools for analyzing prosody (Avanzi et al., 2010), (Goldman et al., 2007) have been developed but tested so far mostly on monologue data. The determination of the relevant units of the different linguistic domains is a crucial issue for this kind of work. In this poster, we will present a series of quantitative evaluations of the output of various automatic tools dealing with prosody, syntax and discourse.

The data we are using is the Corpus of Interactional Data. This is a corpus made of 8 conversations of one hour involving two speakers.

2 Automatic tools

We ran Analor and Momel-Intsint on our corpus. Moreover, we implemented a version of the Simon and Degand (2011) characterization of the prosodic units by using (as they do) the output of Prosogram (Goldman et al., 2007). In the discussions below, the units of Analor are called *Periods* while the ones of Simon and Degand (2011) are called UIM (Major Intonative Units). We also have Interpausal Units (IPU) as a baseline.[1]

At the syntactic level, at the current stage, we simply used a projection of the punctuation learned on a large balanced corpora.[2] More precisely, from a tagged corpus we have learned where strong (periods, exclamation marks etc.) and weak (commas) punctuations occur.

3 Manual Annotations

Among other linguistic elements, prosodic and discourse units have been annotated in the framework of the OTIM project (Blache et al., 2010).

Concerning prosody, several kinds of segmentation have been produced. Originally experts have segmented about 2 hours of corpus into Accentual Phrases (AP) and Intonative Phrases (IP). More recently, an annotation campaign involving naive annotators has been realized. The whole CID corpus has been double-annotated. The task for naive annotators consisted in marking prosodic boundaries of different levels (*1, 2, 3* ; 0 being the default non-annotated case of no boundary).[3]

Concerning discourse, the annotation campaign also involved naive annotators that have segmented the whole corpus (half of it being cross annotated). This was realized thanks to a discourse segmentation manual, inspired by (Afantenos et al., 2010) but largely adapted to our interactional spoken data and simplified to be used by naive annotators. The manual combined semantic (eventualities identification) and discourse (discourse markers)

[1]We do not pretend that these different units are supposed to capture the same prosodic level. We simply want to experiment with the units produced by these tools to decide how to use them later.

[2]This corpus was mostly a written corpus which could be an issue. However, we consider the information captured to be relevant.

[3]This was realized according to a coding manual developed by Roxane Bertrand and Cristel Portes.

		Recall	Precision
IPU	Start	0.415	0.838
	End	0.376	0.736
	Units	0.177	0.245
Period	Start	0.353	0.843
	End	0.339	0.783
	Units	0.153	0.364
UIM	Start	0.478	0.794
	End	0.428	0.710
	Units	0.218	0.360

Table 1: Precision and recall. Reference segmentation: manual IP

	Reference segmentation		
	IP	Discourse	Pseudosyntax
IP	-	0.322	0.603
Discourse	0.238	-	0.435
Pseudosyntax	0.369	0.364	-

Table 2: Divergence between linguistic domains

and pragmatic (recognition of specific speech acts) instructions to create the segmentation. Such a mixture of levels has been made necessary by the nature of the data featuring both rather monologic narrative sequences and highly interactional ones. Manual discourse segmentation with our guidelines has proven to be reliable with κ-scores ranging between 0.8 and 0.85.

4 Experiments

Prosodic units Concerning prosodic units, we first compare the automatic tools to the manual annotation into Intonative Units (See Fig. 1). Overall, IP manual annotation is much more fine grained than the segmentation performed by the automatic tools. It is therefore no surprise to find that precision is rather good (at least for boundary detection) while recall is extremely low. It is noticeable to remark how low are the scores when we shift our attention to unit determination rather than simple boundary detection. By the time of the conference, we will have also compared all the naive annotations (in terms of strength of frontiers) with the automatic tools.

Interfaces In order to try to shade a new light on the interfaces we attempted rather rough quantitative comparison (using the WindowDiff measure (Pevzner and Hearst, 2002)) of the units from the different linguistic domains (See Fig. 2). This was done by using (i) the expert segmentation into Intonative Units, (ii) the manual discourse segmentation and (iii) the projection of the punctuation for the syntactic level.

c_1 (strong punctuation+period)	0.285
c_3 (Pseudophrase+UIM)	0.264
c_5 (strong punctuation+IPU)	0.241
IPU	0.198
UIM	0.217
Period	0.265
strong punctuation	0.419

Table 3: WindowDiff comparison of segmentations combining prosody and pseudosyntax, reference: manual discourse units

Finally, we evaluated automatic tools against the manual discourse segmentation (See Fig. 3). The results is that the IPU baseline provides the closest segmentation to the one of the naive annotators. Quite depressingly, both more sophisticated tools are less related to the manual annotation, adding syntactic information create significant divergences with the manual annotations. Lack of space prevent us from a deeper analysis of this results but the poster will be focused on explaining them and finding solutions.

References

Afantenos, S., Asher, N., et al. (2010). Discourse segmentation manual. Technical report, Toulouse University.

Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2010). A corpus-based learning method for prominence detection in spontaneous speech. In *Speech Prosody 2010-Fifth International Conference*.

Blache, P., Bertrand, R., et al. (2010). Multimodal annotation of conversational data. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 186–191, Uppsala, Sweden.

Goldman, J., Avanzi, M., et al. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken french. In *Eighth Annual Conference of the International Speech Communication Association*.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Simon, A. C. and Degand, L. (2011). L'analyse en unités discursives de base: pourquoi et comment? *Langue francaise*, (2):45–59.

Towards Semantic Parsing in Dynamic Domains

Kyle Richardson and **Jonas Kuhn**
Institute for Natural Language Processing
University of Stuttgart
{kyle,jonas}@ims.uni-stuttgart.de

1 Overview

We describe ongoing work in the area of semantic parsing, which is an emerging subfield in NLP that concerns the task of mapping sentences to formal semantic representations. Recent work in this area has focused on using data-driven methods for learning this mapping, both in a supervised setting and in more complex ambiguous learning settings [Mooney, 08]. In the latter learning scenarios, training examples might be given with several possible target semantic representations, the bulk of which don't relate directly to the particular sentence but are instead part of a broader grounded perceptual context. In such a setting, the aim is to model language as being 'situated' in a potentially wide range of observable events.

Well known work by [Chen et al. 2008] on the sportscaster corpus looks at interpreting soccer commentary in ambiguous contexts where several closely occurring (grounded) events are taking place. For example, a naive language observer might hear commentary such as *The purple goalie kicks out to purple3* in the context of several different actions, and at first be uncertain about which event in view is being described. They present a novel bottom-up learning method for accurately parsing unseen game commentary to symbolic semantic representations by 'observing' ambiguous training games, which has inspired a number of subsequent learning studies.

As pointed out by [Bordes et al. 2010], however, the sportscaster corpus has many shortcomings, most notably its lack of lexical ambiguity and small size. *Contexts* are limited to information about events occurring within a very crude window of time around each comment. In a dialogue setting having more background information might be essential. For example, knowing the referent of 'he' in the sentence *he is cooking in the kitchen* requires having knowledge of which individuals are in the kitchen at this time. Similarly, such contextual information is useful for detecting and learning inferential patterns in language.

The *Grounded World* corpus described in [Bordes et al. 2010] gets at some of the issues, and is a set of English descriptions situated within a virtual house. Sentences in the corpus are often ambiguous and employ pronouns, which must be resolved using information about the state of the house (e.g. the location of objects). The corpus, however, was designed largely for doing named entity recognition, and learning is done in a supervised fashion. We describe an extension to this corpus that looks at learning to interpret these descriptions in an ambiguous learning setting.

2 Grounded World*

Utterance: *while he is sleeping in the bedroom*
Original Annotation: - <friend> - <sleep> - - <bedroom>
Observables*: bring(friend,water,toLoc(bedroom))
get(baby,videogame) **sleep(friend,loc(bedroom))**
World State:
location:**bedroom**<'bed', 'closet', **'friend'**, ...>
location:kitchen<'baby', 'fridge', 'cat',.....>

Figure 1: training example from Grounded World*

The original Grounded World corpus consists of 50k (automatically generated) training sentences, paired with a target set of named entities and a world state description, and 30k testing examples. Inside the simulated house is a fixed set of objects, including, for example, a set of *actors* (e.g. 'father', 'brother'), and a set of *furniture pieces* (e.g. 'couch', 'table'). There is also a fixed set of 15 events, such as *eating* , *bringing*, and *drinking*. For our study, we used a small subset of 7k examples from the training set, and modified the sentences to have syntactic alternations and paraphrases not seen in the initial corpus. The original annotations were expanded to normalized semantic representations, and using the world state information we produced a set of distractor events (the *observables*) intended to represent the background knowledge or uncertainty an observer might have about related or simultane-

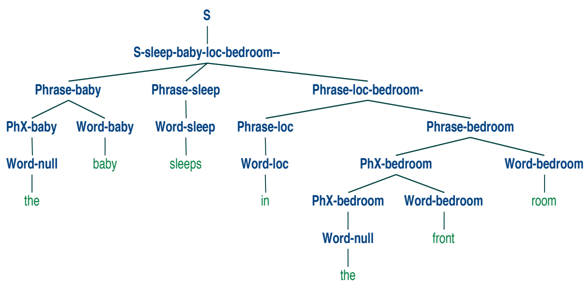


Figure 2: Grounded World* example parse

ous events in view. Figure 1 shows a training example in our Grounded World* corpus, alongside the original annotation. The utterance is *situated* is three separate observable events, two of which are contradictory and represent an observer's uncertainty about whether the friend is involved in the *sleeping* or *bringing* event.

Expanding the relations from the original corpus and *situating* them within larger ambiguous contexts makes the learning task much harder. Given a set of training examples in this narrow domain, we aim to learn, merely from ambiguous observation, how to map novel sentences about the house to their correct semantic representations.

3 Learning

One trend in Semantic Parsing has been to use learning methods that assign rich structure to the target semantic representations, which can be used for finding alignments with latent structures in the language. In many available datasets, the target semantic representations have corresponding *semantic grammars* that produce tree representations. Using ideas from [Wong 2007], [Chen 2008] uses statistical alignment-models for finding alignments between production rules in the semantic grammars and the corresponding words or phrases in the language. In a similar spirit, [Borschinger et al. 2010] recasts the problem in terms of an unsupervised PCFG induction problem, and he develops a technique for automatically generating large PCFGs from the semantic relations in the sportscaster data. In such a setting, the target semantic relations are the *S-Nodes* in the grammar, and the arguments of the relations and relation names are the constituent phrases (in all possible orders) consisting of pre-terminals that correspond to domain concepts. Words in the training data are uniformly assigned to all pre-terminals and the PCFG weights are learned using EM training and the ambiguous contexts as filters.

4 Experiments

In a pilot study to test our extension to the corpus, we adopt the grammar induction technique used in [Borschinger et al. 2010]. We automatically generate a large PCFG using the total semantic relations in our dataset which includes information about the ambiguous contexts. Following the experimental design in [Chen et al. 2008] and [Borschinger et al. 2010], we perform cross validation by making 4 splits in our 7k sentence set (5k for training, and 2k for testing). We then train on each set using the Inside-Out Algorithm, and evaluate by parsing the remaining unseen sentences and compare each S-node relation to a gold standard.[1] In our initial experiments, we don't consider the world state information, and instead resolve pronouns by choosing the most probable analysis observed in the training.

An example analysis produced after training is provided in figure 2, where the derived S-node relation is *sleep(baby, loc(bedroom))*. In the initial experiments, we achieve an average precision of 77.6 % over the four splits. Most errors relate to pronoun resolution, which had an average accuracy of 37.4%. Further work will look at building a parser that considers world information, building on insights from [Schuler 2001].

References

Bordes, A. et al. 2010 Towards Understanding Situated Natural Language. in *Proc. of the 13th AISTATS*

Borschinger, B. et al. 2011 Reducing Grounded Learning Tasks to Grammatical Inference in *Proc. of EMNLP 2011*

Chen, D. and R. Mooney 2008. Learning to Sportscase: A Test of Grounded Language Acquisition. in *Proc. of ICML*

Mooney, R 2007. Learning for Semantic Parsing. in *Proc. of the 8th CICLing*

Schuler, W. 2001. Computational Properties of environment-based disambiguation . in *Proc. ACL*

Wong, Y. et al. 2008. Learning for Semantic Parsing with Statistical Machine Translation. in *Proc. of HLT/NAACL*

[1] we used Mark Johnson's CKY and Inside-Out implementation available at http://web.science.mq.edu.au/ mjohnson/Software.htm.

Why do we overspecify in dialogue?
An experiment on L2 lexical acquisition

Alexandra Vorobyova
École Normale Supérieure
Paris, France
alexandra.vorobyova@gmail.com

Luciana Benotti
PLN Group, FAMAF
National University of Cordoba
Cordoba, Argentina
luciana.benotti@gmail.com

Frédéric Landragin
CNRS, LaTTICe Laboratory
Montrouge and Paris, France
frederic.landragin@ens.fr

Abstract

We describe an experimental study designed to evaluate the effect of overspecification on L2 language acquisition. Our hypothesis is that overspecification helps establish alignment, which facilitates the acquisition of lexemes. Our results show that subjects receiving overspecified references during the exercise phase are slower at resolving the exercises, but have better lexeme acquisition rates. This supports the claim that overspecification is a useful mechanism for communication.

1 Introduction

The study of the generation and interpretation of *referring expressions* (REs) has been an active area of research for many years, due to its importance for communication. In particular, the role of *overspecification* in reference has received much attention – studies have shown that, although it makes interpretation more costly, redundant information is frequently used in the referring expressions produced by speakers and that while subjects take longer to resolve overspecified references, they do not rate them them any worse than minimal ones (Maes et al., 2004; Engelhardt et al., 2006). There are two main competing explanations that have been proposed for the overspecification phenomenon. One explanation holds that (1) overspecification is a result of speakers' cognitive limitations and impairs the comprehension of the REs (Engelhardt et al., 2011). The second claims that (2) overspecification is a useful part of communication because it gives the listener more chances to align with the speaker, compensates for perceptual difficulties, and makes future communication more effective (Nadig and Sedivy, 2002).

In this paper, we aim to support explanation (2) by empirically evaluating the effect of overspecification on lexical acquisition in second language (L2) learning. Our hypothesis is that overspecification helps establish *alignment* (Brennan and Clark, 1996) between the speaker and the listener, which in turn facilitates lexical acquisition.

2 The Experiment

To test our hypothesis, we created an instruction-giving system that produces minimal and overspecified REs of objects located in the context of a 3D virtual world, designed using the GIVE platform (Koller et al., 2008).

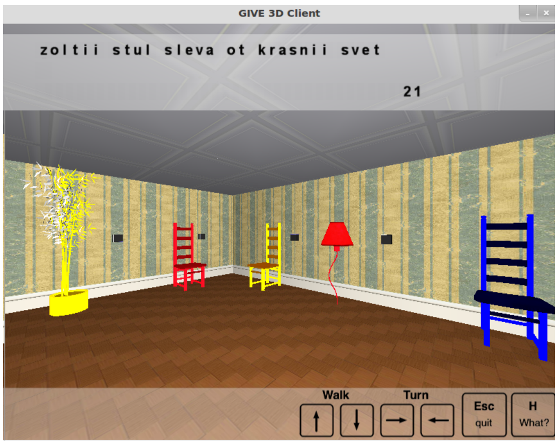


Figure 1: Exercise Phase: Referring expression received by a subject in the OR condition *zoltii stul sleva ot krasnii svet* means 'yellow chair on the left of the red light'.

We recruited fifty subjects and made two equal groups: the MR (Minimal Reference) group received minimal REs regarding objects in the Exercise Room, whereas the OR (Overspecified Reference) group received overspecified REs. Figure 1 shows a RE as received by a subject in the OR condition — REs were overspecified with a relation to a neighbouring object since there are case studies that show that this is the preferred property that is most frequently overspecified in corpora (Viethen and Dale, 2008).

3 Results

In order to test our hypothesis, we extracted information on *whether* and *how much* the number of errors decreased between the First Test Phase and the Second Test Phase of the experiment. In Table 1, we can see that 33% more OR subjects decreased their errors compared to MR subjects, which is represented by the *lexeme acquisition rate*, and that a bigger percentage of errors was overcome in the OR condition (43%) than in the MR condition (29%) (the *error overcoming rate*). We can also see that the *average resolution speed* with which the subjects in each condition resolved the referring expressions in the exercise phase is two times slower for the OR condition than the MR condition. Finally, in a post-experiment questionnaire, we found that OR subjects did not rate the received expressions worse and evaluated that the Exercise Phase as more useful to acquire the lexemes than the subjects in the MR condition.

Metric	MR	OR
Lexeme acquisition rate (%)	56	89
Error overcoming (%)	29	43
Resolution speed (cm/s)[a]	101.1	49.88

[a]The metric unit used for speed is an interpretation of perceived size in the virtual world.

Table 1: Objective metrics gathered during the experiment for the two group of subjects.

Our hypothesis was confirmed by our results: the overall OR lexeme acquisition rate was significantly higher than that of the MR condition and subjects perceived the training exercises as more effective when overspecified REs were used. These results are coherent with previous work that reports that it takes more time to resolve overspecified referring expressions and that overspecified REs are evaluated as equal to minimal ones (Engelhardt et al., 2006).

4 Conclusion

In this paper, we have shown that subjects learning Russian words via a virtual-world task had better success rates when they were provided with overspecified training exercises, and evaluated the exercises as more useful. This has applications in dialogue system development — if overspecification is useful for establishing alignment, then algorithms should produce overspecified references to facilitate communication.

References

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

Paul Engelhardt, Karl Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573.

Paul Engelhardt, Baris Demiral, and Fernanda Ferreira. 2011. Over-specified referring expressions impair comprehension: An erp study. *Brain and Cognition*, 77(2):304 – 314.

Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johana Moore, and John Oberlander. 2008. Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *International Conference in Natural Language Generation*, pages 243–250. Association for Computational Linguistics.

Alfons Maes, Anja Arts, and Leo Noordman. 2004. Reference management in instructive discourse. *Communication And Cognition*, 37(2):117–144.

Aparna Nadig and Julie Sedivy. 2002. Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13(4):329–336.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 59–67, Stroudsburg, PA, USA. Association for Computational Linguistics.