Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL)

Jonathan Ginzburg, Pat Healey and Yo Sato (Eds.)

2-4 June, 2008
London, U.K.
Sponsored by Dialogue Matters

Preface

Contents

Part I

Conference Papers

Extended Abstract: Computational Models of Non-cooperative dialogue

David R Traum Institute for Creative Technologies
University of Southern California
Marina Del Rey, CA 90292 USA
traum@ict.usc.edu

Cooperativity is usually seen as a central concept in the pragmatics of dialogue. There are a number of accounts of dialogue performance and interpretation that require some notion of cooperation or collaboration as part of the explanatory mechanism of communication. For instance, Grice's cooperativity principle and associated maxims are used to explain conversational implicature (Grice, 1975). Searle uses general principles of cooperative conversation to account for indirect speech acts (Searle, 1975). Clark and Wilkes-Gibbs use a principle of "least collaborative effort" as a goal of the processes of grounding and accepting referring expressions. (Clark and Wilkes-Gibbs, 1986).

Alwood (Allwood, 1976) considers that full-blown communication requires at least some degree of cooperation, and defines ideal cooperation between a number of interacting normal rational agents as adherence to the following principles:

1. they are voluntarily striving to achieve the same purposes,
2. they are ethically and cognitively considering each other in trying to achieve these purposes.
3. they trust each other to act according to 1 and 2 unless they give each other explicit notice that they are not'. consisting of four parts:

Most advanced computational work on dialogue agents has also generally assumed cooperativity. Simple dialogue systems, e.g (Sutton et al., 1996), are programmed to react directly to specific types of inputs, without doing much pragmatic reasoning. Some advanced systems are formulated as agents that reason about attitudes such as belief, desire, and intention, e. g. (Cohen and Perrault, 1979; Allen and Perrault, 1980). These systems do means-ends reasoning to develop plans that further their goals which can be adopted as intentions, and also recognize the plans of others. There is still a tension between the model of individual agency and coordinated action, which is often modelled using principles of cooperativity, collaboration including such notions as joint intentions (Cohen and Levesque, 1991) and Shared Plans (Grosz and Sidner, 1990). These notions are used to automate the kinds of pragmatic reasoning described by Grice and Searle and compute speaker meaning using contextual knowledge as well as compositional semantics. This notion of cooperativity in conjunction with rational agency can be a powerful mechanism for allowing systems to engage in human-like flexible dialogues.

The cooperative principles are reasonable for the vast amount of domains that people have built dialogue systems for: service or information providing systems, in which the goals of both the system and user can be seen to coincide. What happens, though, when there is no shared goal, or cooperation breaks down in other ways, e.g., lack of cognitive or ethical consideration (and/or follow-through) or lack of trust? Many models have little to say about this kind of dialogue, and in fact disparage non-cooperative behavior in human-machine dialogue because it "easily leads to miscommunication and an unnecessarily long, complicated, and perhaps failed dialogue because of the system's limited abilities to detect, handle, and recover from a

non-cooperative dialogue flow." (Klein et al., 1999; Hajdinjak and Mihelic, 2004). While cases of dialogue systems that are intentionally non-cooperative are not yet common for human-computer dialogue, there are a number of applications in which they are important, including:

- intelligent tutoring systems, e.g. (Zinn et al., 2002), in which the tutor must sometimes override the local desires of the student for the assumed greater good of education
- commercial bargaining agents, e.g., (Jameson et al., 1994; Jameson and Weis, 1995), in which the buyer and seller have opposite goals, at least in terms of price.
- more generally, assistant agents in which the agent may talk to someone other than its owner, in which it should take up the goals of its owner rather than others whom it may engage in dialogue with.
- role-playing training agents, in which the agents are playing roles which are not cooperative in order to let a user practice and learn how best to deal with such situations. (Traum et al., 2008; Traum et al., 2007)

In (Traum and Allen, 1994), we presented a model of coordinated dialogue behavior that did not rely on cooperativity for basic interaction. In that view, dialogue behavior could be motivated either by individual goals (which might or might not be shared or cooperatively adopted) or obligations, which are imposed by norms of social interaction and can be ignored only with potential social penalties. This had the potential to handle question-answering in non-cooperative situations, but the Trains system which used it (Allen et al., 1995) was highly cooperative.

More recently, we have been working on a number of virtual humans, who engage in face to face spoken dialogue and act as role-players for domains such as non-team negotiation (Traum et al., 2005), as shown in Figure 1, multiparty negotiation, as shown in Figure 2, and questioning interviews (Traum et al., 2007), as shown in Figure 3. In these domains, cooperativity is an achievement rather than an assumption. The agents can choose to be cooperative or uncooperative. Dialogue must proceed in

Figure 1: SASO-ST Negotiation in the Clinic: Dr Perez

both of these cases, and in fact, dialogue is one of the principal means of increasing cooperativity. We thus need accounts of aspects of dialogue behavior in which cooperativity does not play an essential role, as well as other computational mechanisms for specific uncooperative behaviors.

Figure 2: SASO-EN Negotiation in the Cafe: Dr Perez (left) looking at Elder al-Hassan

This talk will outline some cases of noncooperative communication behavior and computational dialogue mechanisms that can support these kinds of behavior, including generating, understanding, and deciding on strategies of when to engage in uncooperative behavios. Behaviors of interest include

- unilateral topic shifts or topic maintenance
- avoidance
- competition
- unhelpful criticism

- withholding of information or services
- lying & deception
- competition
- antagonism
- rejection of empathy

Figure 3: Tactical Questioning: Hassan

The decision of whether to be cooperative or not and how to behave in each case depends on a number of factors, including the standard notions of belief, desire, intention, obligation, and initiative, but also factors such as trust, solidarity, power, status, and respect.

We will present preliminary computational models of these factors and illustrate their use with examples of interactions with the characters shown in Figures 1, 2, and 3.

Acknowledgments

We would like to thank the rest of the Virtual Human team at USC. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

James F. Allen and C. Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178.

James. F. Allen, L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio, and D. R. Traum. 1995. The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 7:7–48.

Jens Allwood. 1976. *Linguistic Communication as Action and Cooperation*. Ph.D. thesis, Göteborg University, Department of Linguistics.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39. Also appears as Chapter 4 in (Clark, 1992).

Herbert H. Clark. 1992. *Arenas of Language Use*. University of Chicago Press.

Phillip R. Cohen and Hector J. Levesque. 1991. Teamwork. *Nous*, 35.

Phillip R. Cohen and C. R. Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212.

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

Barbara J. Grosz and Candace L. Sidner. 1990. Plans for discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press.

Melita Hajdinjak and France Mihelic. 2004. Information-providing dialogue management. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *TSD*, volume 3206 of *Lecture Notes in Computer Science*, pages 595–602. Springer.

Anthony Jameson and Thomas Weis. 1995. How to juggle discourse obligations. In *Proceedings of the symposium on Conceptual and Semantic Knowledge in Language Generation*, Nov.

Anthony Jameson, Bernhard Kipper, Alassane Ndiaye, Ralph Schäfer, Joep Simons, Thomas Weis, and Detlev Zimmermann. 1994. Cooperating to be Noncooperative: The Dialog System PRACMA. In B. Nebel and L. Dreschler-Fischer, editors, *Proceedings of the Eighteenth German Conference on Artificial Intelligence (KI-94)*, pages 106–117. Springer.

Marion Klein, Niels Ole Bernsen, Sarah Davies, Laila Dybkjaer, Juanma Garrido, Henrik Kasch, Andreas Mengel, Vito Pirrelli, Massimo Poesio, Silvia Quazza, and Claudia Soria. 1999. Supported coding schemes.

Deliverable D1.1, MATE Project. available from http://www.dfki.de/mate/d11/.

John R. Searle. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 59–82. Academic Press.

S. Sutton, D. G. Novick, R. A. Cole, and M. Fanty. 1996. Building 10,000 spoken-dialogue systems. In Proceedings 4th International Conference on Spoken Language Processing (ICSLP-96).

David R. Traum and James F. Allen. 1994. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

David Traum, William Swartout, Stacy Marsella, and Jonathan Gratch. 2005. Fight, flight, or negotiate: Believable strategies for conversing under crisis. In *In proceedings of the Intelligent Virtual Agents Conference (IVA)*, pages 52–64. Springer-Verlag Lecture Notes in Computer Science, September.

David Traum, Antonio Roque, Anton Leuski, Panayiotis Georgiou, Jillian Gerten, Bilyana Martinovski, Shrikanth Narayanan, Susan Robinson, and Ashish Vaswani. 2007. Hassan: A virtual human for tactical questioning. In *The 8th SIGdial Workshop on Discourse and Dialogue*.

David Traum, William Swartout, Jonathan Gratch, and Stacy Marsella. 2008. A virtual human dialogue model for non-team interaction. In Laila Dybkjaer and Wolfgang Minker, editors, *Recent Trends in Discourse and Dialogue*. Springer.

Claus Zinn, Johanna D. Moore, and Mark G. Core. 2002. A 3-tier planning architecture for managing tutorial dialogue. In Stefano A. Cerri, Guy Gouardères, and Fábio Paraguaçu, editors, *Intelligent Tutoring Systems*, volume 2363 of *Lecture Notes in Computer Science*, pages 574–584. Springer.

A Grounding Approach to Modelling Tutorial Dialogue Structures

Mark Buckley and **Magdalena Wolska**
Dept. of Computational Linguistics
Saarland University
66041 Saarbrücken, Germany
{buckley|magda}@coli.uni-sb.de

Abstract

Pedagogically motivated analyses of tutorial dialogue have identified recurring local sequences of exchanges which we propose to be analysed analogously to grounding structures. In this paper, we present a model describing such local structures in which a learner and a tutor collaboratively contribute to building a solution to a task. Such structures are modelled as "grounding" exchanges which operate at the task level, i.e. at the level of deep understanding of the domain. Grounding a learner's contributions depends on the tutor's beliefs as to the learner's level of understanding. We treat this explicitly by requiring sufficient domain-level evidence to be shown for a contribution to be grounded. This work attempts to link general theories of dialogue with observations from pedagogical science.

1 Motivation

Successful conversational communication depends strongly on the coordination of meanings and background assumptions as to the state of the world (Clark, 1992; Stalnaker, 2002; Thomason et al., 2006). Dialogue participants try to achieve a situation in which they mutually believe that their utterances are interpreted as intended and that their assumptions as to the shared knowledge, the *common ground*, agree. To this end, they engage in a process called *grounding* (Clark and Schaefer, 1989; Traum, 1999), whose purpose is to ensure explicit alignment of (mutual) beliefs. Grounding can serve to avoid or recover from communication failures arising from problems which may range from low level signal-related issues through the interpretation of the propositional content up to the level the communicative intentions of speech acts.

Grounding is a general pragmatic phenomenon in cooperative communication that is independent of the purpose of the verbal activity, be it socially-motivated spontaneous conversation or task oriented verbal communication such as information seeking, negotiation, problem solving or dialogue-based instruction. The latter scenario is additionally inherently prone to misalignment of beliefs beyond the level of the communicative intentions of speech acts: namely at the level of *deep understanding* of the tutored domain. First, tutoring is typically characterised by an asymmetry of knowledge possessed by the tutor and the learner (Munger, 1996; Lee and Sherin, 2004). Second, there is an uncertainty on the part of the tutor as to the learner's deep understanding and the overall knowledge state. In fact, empirical research shows that tutors tend to have difficulties in estimating the learner's deep understanding (Chi et al., 2004). Still, dialogue-based one-on-one instruction, even by non-experts, has been shown to produce higher learning gains than other forms of teaching (Bloom, 1984; Moore, 1993). One of the factors that makes a difference in the efficiency of instruction is adaptivity of tutorial feedback and explanation. Nückles et al. (2006) show that tutors who are better informed on the learners' prior knowledge can better adapt their feedback. Another important feature of efficient tutoring are locally targeted pedagogical actions. Graesser et al. (1995) show in an empirical study that tutors typically do not focus on cognitive alignment, i.e. do not strive to establish complete understanding of the students' state of beliefs. Instead they tend to perform specific targeted tutoring moves that locally address

the student's (lack of) progress on the task at hand.

Motivated by these findings we have been investigating discourse and dialogue phenomena in the context of dialogue-based tutoring with the ultimate goal of building a tutoring system for mathematical theorem proving. Our approach to modelling tutorial dialogue draws on the empirical evidence from the above-mentioned studies and can be summarised by the following observations:

On the one hand, tutorial dialogue is in many respects different from other types of dialogue. The model of cooperative interpretation must address the learner's utterances not only in terms of their function as speech acts, but also as demonstrations of knowledge, that is, it is dependent on the adopted pedagogical strategy. Motivated by pedagogical goals and licenced by his authority, the tutor may be the "uncooperative" interlocutor in the sense that he/she may demand presentation of pieces of knowledge that the learner had left implicit, or may even refuse to provide information requested by the learner (overriding dialogue obligations valid in other dialogue genres) attempting to lead the learner to self-discovery of knowledge. The structure of tutorial dialogue is moreover characterised by systematically recurring sub-structures. The role of these is to address the learner's knowledge contributions and to monitor, at least to some extent, the learner's deep understanding, allowing feedback and cooperative behaviour to be adapted to what the student has previously shown to have understood.

On the other hand, tutorial dialogue is still a type of dialogue, that is, it is characterised by the general phenomena present in any dialogue genre and should lend itself to modelling in terms of general notions of dialogue. However, because it is a special type of dialogue, the model's parameters (e.g. the contents of the information state, models of dialogue state transitions, obligations, and cooperativity) must be adjusted to the genre's characteristics.

This work is an attempt to apply notions from general dialogue theory to tutorial dialogue. In particular, we will try to show parallels between the structure of grounding at the speech acts level and the local structures in tutorial dialogue which resemble grounding, but address the deep understanding of the domain. We will call these *communication level grounding* and *task level grounding* respectively. We start by exemplifying these local structures with dialogue excerpts from our corpora (Section 2.1). In Section 2.2 we briefly introduce grounding according to Traum (1999). In Section 2.3 we present our framework for task level grounding and point at the differences between it and Traum's model. Section 3 presents the model formally and steps through an example, before we summarise our conclusions in Section 4.

2 Tutorial Dialogue Structures as Grounding Exchanges

Tutorial dialogues have been shown to exhibit local patterns, referred to by Graesser et al. (1995) as *dialogue frames*, related to the pedagogical goals that tutors follow. We will argue that the structure of dialogue frames is similar in character to that of Traum's Discourse Units (Traum, 1999), the basic building blocks of which are utterances which contribute to achieving mutual understanding. Our goal is to attempt to unify these two views on (tutorial) dialogue structure in a grounding-based model of tutorial dialogue, which we present in the next section.

2.1 Dialogue Frames in Tutoring

In a corpus-based analysis of the collaborative nature of tutorial dialogue Graesser et al. identify local interaction patterns which make one-on-one tutoring, even by non-experts, effective in producing learning gains. They consist of the following steps, all of which but step 2 may be omitted:

Step 1 Tutor asks a question.

Step 2 The student offers an answer

Step 3 Tutor gives feedback on the answer

Step 4 Tutor and student collaboratively improve the quality of the answer, whereby the tutor can for instance elaborate on the answer, give a hint, pump the student for more information, or trace an explanation or justification.

Step 5 The tutor assesses the student's understanding of the answer, for instance by explicitly asking whether the student understood.

Similar structures were revealed by our analysis of the two corpora of tutorial dialogues on mathematical theorem proving (Wolska et al., 2004;

Benzmüller et al., 2006) which we have collected.[1] (1) and (2) are examples of such exchanges[2]:

(1)
S0 $(R \circ S)^{-1} = \{(x,y)|(y,x) \in (R \circ S)\}$
T0 correct
S1 $(R \circ S)^{-1} = \{(x,y)|(y,x) \in \{(x,y)|\exists z(z \in M \wedge (x,z) \in R \wedge (z,y) \in S\}\}$
T1-1 okay,
T1-2 but you could have done that more simply

(2)
S19-1: $(R \circ T) = (T^{-1} \circ R^{-1})^{-1}$ (by exercise W),
S19-2: then it must also hold that $(S \circ T) = (T^{-1} \circ S^{-1})^{-1}$
T25: Why does this follow from exercise W?
S20: $(R \circ S) = (S^{-1} \circ R^{-1})^{-1}$ (according to exercise W), then it must also hold that $(S \circ T) = (T^{-1} \circ S^{-1})^{-1}$ and $(R \circ T) = (T^{-1} \circ R^{-1})^{-1}$
T26-1: All other steps are appropriate,
T26-2: but the justification for $(R \circ T) = (T^{-1} \circ R^{-1})^{-1}$ is still missing.
S21: $(R \circ T)^{-1} = (T^{-1} \circ R^{-1})$ (by exercise W)
T27: Yes.

The building block of such exercises is the *proof step*, a contribution which consists of a formula which the step derives, a justification, premises, and possibly other components. Proof steps may be underspecified, for instance by only providing the derived formula. This leads to them possibly having to be augmented in order to be acceptable.

In (1) we see a simple case of a student's contribution being accepted by the tutor. In terms of Graesser's dialogue frames, S0 corresponds to step 2 and T0 to step 3. Because the tutor is immediately satisfied that the student has understood the answer, steps 4 and 5 are not performed. S1 and T1 form a new dialogue frame which is the same as the first except that step 4 is realised in T1-2 by the tutor, who elaborates on the answer.

(2) is a more complex example which begins with the student's contribution in S19 (Graesser's step 2). It consists of two contributions, however only the first one (S19-1) is discussed. Similarly to S0 in (1), the contribution is incomplete in that the student does not provide the premise that allowed him/her to conclude that the contribution in S19-1 holds, but rather leaves it implicit. Here however, the tutor is not satisfied with the incomplete step and responds with a request to elaborate the answer (step 4) in T25, skipping the feedback (step 3). Instead of addressing this request, the student offers a new contribution, leaving the request in T25 pending. In T26-1 the tutor gives feedback on both S19 and S20 (step 3 for both of these contributions) and continues by repeating the request for elaboration in T26-2. The student then addresses this request by supplying the missing premise in S21 (step 4) which the tutor accepts in T27, thereby closing the dialogue frame.

Our analysis of the two tutorial dialogue corpora revealed that structures such as the ones described above systematically recurred in the domain of proof tutoring in the context of the conducted experiment. Locally, the dialogue structures indeed typically reflect Graesser's steps 2 though 4, with individual proof steps being proposed (step 2) and subsequently optionally elaborated (step 3) and evaluated (step 4), in either order. Due to the student having the initiative in our experimental setup step 1 is seldom found in our data.

In the corpora elaboration requests were most commonly initiated because the inferences proposed by the students were only partially specified. Typically, the students provide a formula (or an equivalent worded statement) leaving out, for instance, the inference rule, the way it should be applied or the premises. This means that part(s) of the task-level steps are left implicit (or *tacit*), possibly resulting in them not being grounded. In the tutoring domain the question of whether an underspecified step (or more generally, an incomplete knowledge demonstration) can be accepted (i.e. grounded) depends, for instance, on pedagogical factors (in the case of mathematical proofs, for example on the tutor's notion of an "acceptable" proof (Raman, 2002)) and the tutor's beliefs as to the student's knowledge state.

2.2 The Grounding Acts Model

Traum (1999) defines a set of Grounding Acts which are identified with particular utterance units and perform specific functions towards the achievement of mutual understanding. The content of an utterance can become grounded as a result of an exchange containing Grounding Acts; such possibly multi-turn sequences are referred to as *Discourse Units* (DU). DUs can contain the following acts:

Initiate begins a new DU with a new utterance unit.
Continue adds content to an act.

[1] The corpora were collected in Wizard-of-Oz experiments. The 2004 corpus contains 22 dialogues (775 turns in total) in the domain of naive set theory. The 2006 corpus contains 37 dialogues (1917 turns) in the domain of binary relations.

[2] Sx and Tx label student and tutor utterances respectively.

Acknowledge is evidence of understanding of the current utterance unit. This evidence can be of differing strength, e.g. demonstration of the understood meaning or performance of a relevant continuation.

Repair changes the content of the current utterance unit.

ReqRepair requests repair of a previous act by signalling non-understanding.

ReqAck asks the dialogue participant to acknowledge understanding of a previous act.

Cancel abandons the current DU without grounding.

In the DU in example (3), taken from Traum (1999), the content of the initiating utterance I1-1 and the continuation I1-2 has been successfully grounded by the acknowledgement Grounding Act in R1.

(3)

I1-1 Move the box car to Corning	init^I
I1-2 and load it with oranges	cont^I
R1 ok	ack^R

Our previous example (1) also exhibits this structure within the DU, where S0 and S1 are initiations and T0 and T1-1 are acknowledgements. We now show that in tutorial dialogue, in addition to the communicative level of Traum's model, grounding also operates at the task level.

2.3 Grounding in Tutorial Dialogue

We have exemplified the parallels between the structures found in tutorial dialogue and grounding exchanges and will now make these parallels more explicit. We will interpret dialogue frames as discourse units and the actions within dialogue frames as grounding acts.

What is grounded in the course of a discourse unit is a piece of domain content which contributes to the domain-level task. In tutoring this is a knowledge demonstration — we will use the term *solution step*. Proof steps become grounded by being first proposed and then accepted by the tutor, provided that the tutor has sufficient evidence to believe that the student has deeply understood how the step was derived. To reach this state the student may be obliged to supply evidence of having understood the step, and this evidence can be of varying strength. In this sense supplying evidence is similar to Traum's **Acknowledge**, and a request for evidence is similar to **ReqAck**. We list the set of actions as well as who can perform them in the course of grounding a solution step in Table 1.

Propose	S,T	propose a solution step
ReqEv	S,T	request evidence showing understanding of the current step
SuppEv	S	give evidence showing understanding of the current step
Accept	T	accept that the student has understood the current step
Reject	T	reject the step (due to incorrectness or non-understanding)

Table 1: Task level grounding actions and speakers

Augment	an elaboration of the current step
Reword	paraphrase of the current step
Claim	positive answer to "do you understand?"
Verbatim	repeat back the step verbatim

Table 2: Types of evidence of understanding

In the same way that Clark and Schaefer (1989) identify different types of evidence of understanding, the action **SuppEv** encompasses a number of different ways of showing understanding of a solution step. From our analysis of the data, we propose the four categories listed in Table 2 from strongest to weakest. Although verbatim repetition of the content being grounded is the strongest evidence type in Clark and Schaefer's communication level grounding model, at the task level it is the weakest form, since it does not show any understanding beyond recognition of the original signal. Claiming understanding is self-reflection on the student's own belief state, and for our purposes is a weak form of evidence. Rewording is a strong indication of understanding, but does not add anything to the current content which is being grounded. The strongest evidence type is augmenting the current solution step with further information. This shows that the student understands even those components which were not stated in the proposal phase of the discourse unit. In keeping with Clark and Schaefer's observation that evidence must be "sufficient for the current purpose", the tutor's decision of whether to consider this evidence *sufficient* to show understanding of the current content (and then to accept the step) depends on both a student model and the pedagogical strategy being followed. Indeed for different teaching domains this notion of *sufficient* will be defined differently according to the demands of the task and the domain dependent teaching goals.

According to this model the annotation of example (2) from the previous section, where subscripts index individual steps under discussion, is:

S19-1	**Propose**$_1$
S19-2	**Propose**$_2$
T25	**ReqEv**$_1$
S20	**Propose**$_3$
T26-1	**Accept**$_{2,3}$
T26-2	**ReqEv**$_1$
S21	**SuppEv**$_1$ (Augment: premise)
T27	**Accept**$_1$

The proposal made in S19-1 is eventually grounded in T27, but in between a new proposal is made (S20), showing that more than one solution step can be under discussion at once.

Contrasts with Traum's model We have borrowed many concepts from and shown parallels to Traum's Grounding Acts model, so here it is useful to highlight some key differences. The main difference is that task level grounding works at a higher level than Traum's communication level grounding model. Our model does not deal with meaning but rather with deep understanding, and the object being grounded is part of the task being explicitly talked about. Accordingly, actions contributing to task level grounding are motivated by task level goals, such as completing the current exercise, whereas Traum's Grounding Acts contribute to successful communication as a whole. Communication level grounding does however still operate as usual in parallel. We refer to example (2), in which the utterance T25 has two functions: at the communication level it grounds the propositional content initiated in S19 but at the task level it continues the discourse unit. A further difference is the roles of dialogue participants and their goals. In tutoring our model does not consider the roles of speaker and hearer, but rather student and tutor, necessary because of the asymmetry of roles in tutorial dialogue; students are obliged to demonstrate understanding but tutors are not.

In summary, we have found a correspondence between general grounding structures and the structures found in tutoring. In order to treat these subdialogues in terms of grounding we need a model of grounding with a higher level object: the task level step. In the next section we introduce the more formal machinery to model these sequences.

3 A Model of Task Level Grounding

Our discourse unit is a subdialogue which begins with the proposal of a task level step and which ends with this step being either accepted or rejected by the tutor. In the previous section we have motivated this choice by showing its equivalence to both Graesser's dialogue frames and Traum's Discourse Units. The objects which are under discussion and which are to be grounded in these subdialogues are *solution steps*, here proof steps, and the conditions which affect this are a student model, the tutor's pedagogical strategy, the correctness, relevance and granularity of the step, as well as some definition of what it means for evidence to be sufficient. The internal structure of solution steps should be defined for the task at hand — here we use a solution step for mathematical proofs consisting of a formula which is derived, a justification for deriving the formula, and the premises used by the justification. In this section we present the machinery necessary to model these phenomena and step through example (2).

We assume that the dialogue system has access to two expert systems: a pedagogical manager and a mathematical domain reasoner. The pedagogical manager (Fiedler and Tsovaltzi, 2003) is responsible for the teaching strategy that the system follows, as well as for maintaining the student model. The domain reasoner (Dietrich and Buckley, 2007; Schiller et al., 2007) evaluates solution steps with respect to correctness, granularity and relevance, and can resolve missing components of underspecified steps.

The model uses the categorisations of utterance types in terms of their function in the DU (Table 1) and evidence types (Table 2) that play a role in the grounding exchanges we are considering. We will now additionally define a dialogue state which represents intermediate stages of the discourse unit, followed by a finite state machine which encodes the transitions between dialogue states and their effects.

3.1 Dialogue State

The dialogue state used in our model is an extension of our previous work on common ground (Buckley and Wolska, 2007), reduced to those aspects relevant to this presentation. It consists of four parts and is shown in Figure 1. The common ground (CG)

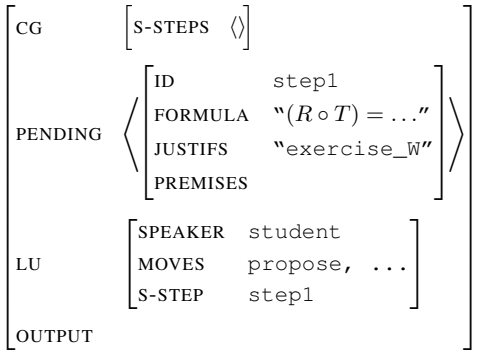


Figure 1: The dialogue state

contains an ordered list[3] of the solution steps which have been grounded in the process of solving the task (S-STEPS). The solution steps which are currently under discussion but are not yet grounded are stored in PENDING. The latest utterance (LU) in the dialogue is represented by a structure containing information about the speaker who performed the utterance, the dialogue moves it realised, and the solution step, if any, that it contained. Finally the dialogue moves that the system should verbalise next are collected in OUTPUT. Both LU/MOVES and OUTPUT store complete dialogue moves, however here we only list task-level grounding actions. When task-level grounding has been successful, the solution step moves from PENDING to CG/S-STEPS.

3.2 Transitions between Dialogue States

Figure 2 presents a finite state machine encoding the transitions between dialogue states in a discourse unit. A **Propose** moves the dialogue into a state in which there is an ungrounded solution step. From here the tutor can either accept the step directly, thus grounding the step, or ask for further evidence of understanding, after which it is necessary for the student to supply evidence before the discourse unit can be in the state in which the solution step is grounded.

The transitions (Table 3) are given as sets of preconditions and effects on the dialogue state. We omit additional processing such as information exchange with system modules. The conditions we use are stated informally — "evidence (in)sufficient" is decided by the pedagogical module, drawing on information from the dialogue state as well as its own

[3]This is a strong simplification — a complete treatment would require a more detailed structure for solution steps.

t1	pre	**Propose** ∈ LU/MOVES
	eff	PENDING := LU/S-STEP
t2	pre	evidence insufficient, ne(PENDING)
	eff	OUTPUT := **ReqEv**
t3	pre	evidence sufficient, ne(PENDING)
	eff	OUTPUT := **Accept**,(feedback)
		push(CG/S-STEPS,pop(PENDING))
t4	pre	**SuppEv** ∈ LU/MOVES
	eff	possibly update solution step
t5	pre	evidence insufficient
	eff	OUTPUT := **ReqEv**
t6	pre	evidence sufficient
	eff	OUTPUT := **Accept**,(feedback)
		push(CG/S-STEPS,pop(PENDING))

Table 3: Preconditions and effects of transitions (ne denotes "non-empty")

student model. Transition t3 moves from a state in which a solution step has been proposed to a state in which that solution step has been grounded. If the evidence for understanding the step is sufficient, and there is content under discussion (ne(PENDING)), then an **Accept** and possibly some feedback is generated, and the solution step is moved from PENDING to CG/S-STEPS. This transition equates to Graesser's step 3 in the dialogue frame. Transitions t2 and t5 both cover the situation where the evidence presented is not sufficient to show understanding, and both result in **ReqEv** being generated, and the solution step(s) that were in PENDING remain there (Graesser's step 4). When evidence is supplied, we follow transition t4, which updates the solution step in the event that evidence of the type **Augment** was supplied. Although it is not included in the FSA, at any stage a discourse unit can be abandoned, possibly with a **Reject** action. This decision can be taken for instance in the state "evidence supplied" when the tutor believes that the student will not be able to show understanding of the step.

Because there can be more than one solution step under discussion at one time, as in example (2), we assume that a separate instance of the FSA is run for each one. An acceptance can thus address more than one solution step. Like downdating questions under discussion, we allow acceptances to ground as many solution steps as necessary. We also note that transitions in the model are only made in reaction to task-level grounding actions, so that as long as other actions are being performed, the FSA stays

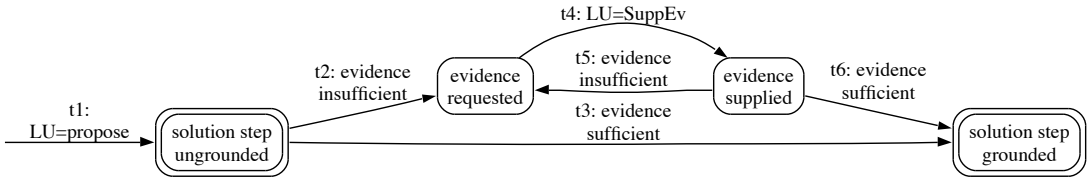


Figure 2: The FSA describing task-level discourse units

in the same state. This allows other levels in the dialogue to be modelled, for instance communication level grounding, off-topic talk or meta talk. Indeed this model can be integrated with a computational model of communication level grounding such as presented by Matheson et al. (2000) if we assume that their grounding acts are dealt with before generating any task level grounding actions. This way problems at the communication level are handled before understanding problems at the task level.

Example Figure 1 shows the dialogue state after utterance S19 in example (2), where the **Propose** in utterance S19-1 has put a solution step in PENDING. The tutor considers that with the current context and student model, there is not sufficient evidence of understanding of the solution step. Transition t2 is therefore executed, generating a **ReqEv** action, realised in utterance T25. Skipping forward to S21 (S20 to T26-2 deal similarly with a different solution step), we recognise a **SuppEv** action, which takes us through transition t4. Since the evidence supplied in S21 is of type **Augment**, we update the solution step by adding the premise the student stated as shown:

$$\left[\text{PENDING} \left\langle \begin{bmatrix} \text{ID} & \texttt{st1} \\ \text{FORMULA} & \text{“}(R \circ T) = \ldots\text{”} \\ \text{JUSTIFS} & \text{“}\texttt{exercise\_W}\text{”} \\ \text{PREMISES} & \text{“}(R \circ T)^{-1} = \ldots\text{”} \end{bmatrix} \right\rangle\right]$$

Now the tutor can reassess whether this more complete solution step is evidence that the student has understood fully, and finds that it is. The transition t6 then generates the **Accept** in T27 and additionally moves the solution step to the common ground. The final dialogue state is shown in Figure 3.

$$\left[\begin{array}{ll} \text{CG} \mid \text{S-STEPS} & \left\langle \begin{bmatrix} \text{ID} & \texttt{st1} \\ \text{FORMULA} & \text{“}(R \circ T) = \ldots\text{”} \\ \text{JUSTIFS} & \text{“}\texttt{exercise\_W}\text{”} \\ \text{PREMISES} & \text{“}(R \circ T)^{-1} = \ldots\text{”} \end{bmatrix} \right\rangle \\ \text{PENDING} & \langle\rangle \\ \text{LU} & \begin{bmatrix} \text{SPEAKER} & \texttt{tutor} \\ \text{MOVES} & \texttt{accept, ...} \\ \text{S-STEP} & \end{bmatrix} \\ \text{OUTPUT} & \end{array}\right]$$

Figure 3: Final dialogue state

4 Conclusions and Related Work

We take advantage of observations about recurring local structures in tutorial dialogue highlighted by Graesser's analysis and recognise that there exist systematic parallels between these and Traum's grounding exchanges. This motivates our computational model, which is analogous to Traum's but operates on a level which directly addresses understanding of the domain. Our model sees these local structures as discourse units whose objects are solution steps, and thus operates at the task level. It captures learners' deep understanding of the domain, and so acts higher than the communication level.

Grounding serves to build up a model of interlocutors' belief states. In tutoring this is particularly important because the tutor's model of the student's belief state is a parameter which affects the adopted pedagogical strategy. The local dialogue structure that our model describes allows the pedagogical model to elicit evidence of understanding and thus reach conclusions about the student's belief state. While we do not make any claims about how such a student model should be constructed, our model does provide input for the construction of a representation of the student's knowledge.

Rickel et al. (2002) also use a general dialogue model in a tutoring system which combines pedagogical expertise with collaborative discourse theory and plan recognition. Their approach models the knowledge state based on steps that the student has been exposed to, however without consid-

ering whether these were fully understood. Zinn et al. (2005) present a tutorial dialogue system which maintains common ground in the dialogue model, however they do not make use of grounding status to structure the dialogue locally. Baker et al. (1999) highlight the necessity for communication level grounding in collaborative learning, but admit that this does not guarantee "deeper" understanding. In general task-oriented dialogues Litman and Allen (1987) derive the structure of clarification subdialogues based on task plans and the discourse structure. Our approach is conceptually similar, however our task model is maintained externally to the dialogue model. Finally, our work relates to that of Thomason et al. (2006) and Benotti (2007) in the sense that the task level grounding model attempts to ground objects that can be viewed as tacit actions.

Our future work will include extending the model to allow more student initiative, for example in the case of domain level clarification requests by the student, as well as looking into more fine-grained structures within the common ground, for instance to support a model of the salience of task level objects.

References

M. Baker, T. Hansen, R. Joiner, and D. Traum. 1999. The role of grounding in collaborative learning tasks. In *Collaborative Learning. Cognitive and computational approaches*, pages 31–63. Pergamon, Amsterdam.

L. Benotti. 2007. Incomplete knowledge and tacit action: Enlightened update in a dialogue game. In *Proc. of DECALOG-07*, Rovereto, Italy.

C. Benzmüller, H. Horacek, H. Lesourd, I. Kruijff-Korbayová, M. Schiller, and M. Wolska. 2006. A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material. In *Proc. of LREC-06*, pages 1766–1769, Genoa, Italy.

B. Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13(6):4–16.

M. Buckley and M. Wolska. 2007. Towards Modelling and Using Common Ground in Tutorial Dialogue. In *Proc. of DECALOG-07*, pages 41–48, Rovereto, Italy.

M. T. H. Chi, S. A. Siler, and H. Jeong. 2004. Can Tutors Monitor Students' Understanding Accurately? *Cognition and Instruction*, 22(3):363–387.

H. H. Clark and E. F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

H. H. Clark, editor. 1992. *Arenas of Language Use*. University of Chicago Press and CSLI.

D. Dietrich and M. Buckley. 2007. Verification of Proof Steps for Tutoring Mathematical Proofs. In *Proc. of AIED-07*, pages 560–562, Los Angeles, USA.

A. Fiedler and D. Tsovaltzi. 2003. Automating Hinting in Mathematical Tutorial Dialogue. In *Proc. of the EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management*.

A. C. Graesser, N. Person, and J. Magliano. 1995. Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, 9:495–522.

V. R. Lee and B. L. Sherin. 2004. What makes teaching special? In *Proc. of ICLS-04*, pages 302–309.

Diane J. Litman and James F. Allen. 1987. A Plan Recognition Model for Subdialogues in Conversation. *Cognitive Science*, 11(2):163–200.

C. Matheson, M. Poesio, and D. Traum. 2000. Modelling grounding and discourse obligations using update rules. In *Proc. of NAACL-00*, pages 1–8.

J. Moore. 1993. What makes human explanations effective? In *Proc. of the 15th Meeting of the Cognitive Science Society*, pages 131–136, Hillsdale, NJ.

R. H. Munger. 1996. Asymmetries of knowledge: What tutor-student interactions tell us about expertise. Annual Meeting of the Conference on College Composition and Communication, Milwaukee, WI.

M. Nückles, J. Wittwer, and A. Renkl. 2006. How to make instructional explanations in human tutoring more effective. In *Proc. of the 28th Annual Conference of the Cognitive Science Society*, pages 633–638.

M. J. Raman. 2002. *Proof and Justification in Collegiate Calculus*. Ph.D. thesis, UC Berkeley.

J. Rickel, N. Lesh, C. Rich, C. Sidner, and A. Gertner. 2002. Collaborative discourse theory as a foundation for tutorial dialogue. In *Proc. of ITS-02*, pages 542–551.

M. Schiller, D. Dietrich, and C. Benzmüller. 2007. Towards computer-assisted proof tutoring. In *Proc. of 1st SCOOP Workshop*, Bremen, Germany.

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

R. Thomason, M. Stone, and D. DeVault. 2006. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In *Presupposition Accommodation*. (draft).

D. Traum. 1999. Computational models of grounding in collaborative systems. In *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 124–131.

M. Wolska, B. Q. Vo, D. Tsovaltzi, I. Kruijff-Korbayova, E. Karagjosova, H. Horacek, M. Gabsdil, A. Fiedler, and C. Benzmüller. 2004. An annotated corpus of tutorial dialogs on mathematical theorem proving. In *Proc. of LREC-04*, pages 1007–1010, Lisbon.

C. Zinn, J. D. Moore, and M. G. Core. 2005. Intelligent information presentation for tutoring systems. In *Intelligent Information Presentation*. Kluwer.

Flexible dialogue management and cost-models

Ian Lewin
University of Cambridge Computer Laboratory
email: ian.lewin@cl.cam.ac.uk

Abstract

Flexibility in dialogue management requires not just the ability to understand and respond to a greater range of user utterance types (or *moves*), but also the ability to generate them and to do so strategically in accordance with some notion of costs and benefits. We explore this issue in the context of the Information State Update model of dialogue. We add costs and preferences to a simple instantiation of the model and explore the added flexibility this brings and also link the inclusion of costs to other developments of the model. We compare this work to the work in reinforcement learning which also includes a notion of cost and reward.

1 Introduction

The Information State Update (ISU) approach to dialogue modelling is a highly abstract characterization of dialogue semantics. Contributions to a dialogue are treated like programs in dynamic logic: they both *update* a dialogue state and are *interpretable* in the light of a previous state.

Agents have two main roles in this abstract picture: to use state in interpreting contributions; and to make state by generating contributions. A great deal of research has concentrated on the former question. What must a state look like if I am to be able to interpret *this* sort of conversational offering? And what will it look like once I have both interpreted and incorporated it? I want to ask: what must a state look like, if I am to choose to make *this* sort of conversational offering? And what will it look like once I have made it? My focus will be on the role of the dialogue state in spoken dialogue systems, partly because this is a useful (and largely externally imposed) constraint on the extent of the material to consider, partly because the results may be practically useful.

The ISU approach to dialogue modelling easily accommodates dialogue models that are encodable directly in a network formalism with atomic states and transitions between them. Choices over dialogue moves can be encoded in a nondeterministic network. The frame based (or slot-and-filler) architecture, for example as instantiated in the form interpretation algorithm in the commercially employed VoiceXML dialogue specification (Oshry et al., 2006) is also easily implementable. An example frame for the travel planning scenario is shown in figure 1. The different instantiations of the frame form the states of the system. Dialogue policy is implemented by associating a question with each attribute in the frame and by ordering the attributes. The next question to be put is the first attribute for which no value is currently known. The whole frame may also be associated with a question, which is put when a frame first comes into focus. The frame stays in focus until all attributes have received values. The VoiceXML question selection algorithm is: if the frame is already in focus, ask the first question (from top to bottom) whose value is unknown else ask the question associated with the whole frame.

The general slot-and-filler approach has continued to underpin a large amount of research in dialogue systems including theoretical work that is implemented in demonstration systems and in systems

that attempt to learn dialogue strategy. In the next section, we review this work and highlight the need for informing machine learning approaches with insights from theory as well as the need for the theoretical approaches to include explicit cost models. Following this, we add a simple cost and preference model to an Information State Update based dialogue manager and explore some of its implications. With a focus on practical systems, we consider the dialogue management of questions that are not directly answered, comparing the treatment with that of models which add features such as "questions under discussion". We conclude that the addition of a cost model is vital not only for future theoretical work but also as a basis for informing future targets for machine learning approaches.

2 Previous Work

Dialogue strategy development in the Information State Update approach has followed two main trends. Theoretical work has concentrated mostly on extending the notion of dialogue state in order to permit analysis of a greater range of dialogues and a deeper analysis of phenomena such as grounding. Secondly, there has been a strand directed towards learning dialogue strategy automatically from real or simulated dialogues. For computational reasons, this work has tended to use a much narrower conception of state. Indeed, the focus of attention has almost entirely been devoted to learning whether and how to confirm user utterances and whether the next question should be more open-ended or not ("mixed initiative"). Although the notion of state in this work tends to be more limited, the strategy is clearly linked to notions of costs and progress towards dialogue goals.

Chu-Carroll and Nickerson (2000) and Litman (2002) describe experiments showing that online adaptation in dialogue strategy is beneficial to users. Litman and Pan's TOOT system monitors predicted speech recognition error rates and can change strategy twice during a dialogue. The system begins by not confirming user utterances at all. If problems are detected, it can start implicitly confirming utterances ("I heard you say Sunday. What time would you like to leave?"); and if problems continue, it can move to explicit confirmation ('On which day of the week do you want to leave", "Sunday", "Do you want to leave on Sunday?"). Simultaneously, the open-endedness of the prompts and the range of acceptable responses is degraded. For example, at the middle stage (entitled *mixed initiative*, somewhat bizarrely) the system will not let the user ignore its question but insists on an answer. Chu-Carroll and Nickerson's MIMIC system also monitors more general interpretation difficulties than just speech recognition problems. A binary global switch can be set causing the system to offer less open-ended prompts. The switch may also be reset if sufficient cue evidence of success can be found. One cue is whether the user subsequently adds unsolicited information to an answer.

Reinforcement learning techniques have also been widely explored as a means of data-driven development of dialogue strategy. In addition to a corpus of dialogues encoding dialogue states and transitions between them, a reward function is required which enables the learning function to generate a strategy that generally leads to higher rewards.

For example, Scheffler and Young (2002) used a corpus of dialogues generated with a user simulation to learn how to confirm what was just said (explicitly, implicitly by repeating whilst querying the next slot or not at all) and whether to offer open-ended questions or not. In addition to slot information, the state contained confidence scores for the latest recognition result. Scheffler and Young argue that it is appropriate to restrict the learning to suitably *local* decision making (such as confirmation of last utterance) on the grounds that the user simulation provides the data for making these choices but not on the higher strategic questions such as which slot to ask for next. In contrast, Henderson (2005) explores the learning of strategies for the entire dialogue and also uses a much richer notion of dialogue state including possibly its entire history. The learned strategies are reported to perform well when tested against user simulations and against real users (Lemon et al., 2006). Qualitative analysis of the dialogues (Frampton and Lemon, 2006) suggests that the improved performance of the learned strategies is actually entirely attributable to improved (local) repair strategies when the last exchange failed to add a value for a slot. Rather than simply repeating a failed question the improved strategies would either

give help or switch focus to a different slot. Giving help is an evidently sensible strategy. The switch of focus to another slot is interesting however not least because the reward function only rewarded dialogues where all slots are filled and the learner did learn only to query the database when all slots had been filled. Consequently, switching focus could only postpone acquisition of a value for a particular slot. The system apparently learned that exact question repeats tend to be unsuccessful. In general, question repeats may be unsuccessful if answer pronunciation remains identical or indeed if the answer becomes hyper-articulated. Possibly these properties were reproduced in the n-gram user simulations used for testing which were trained from Communicator data. In this case, asking the same question again later but not actually immediately might be more likely to result in a recognizable response. However there is clearly a danger that the learned strategies are just responses to somewhat unnatural artifacts of the user simulation. Whether such a strategy could actually be viable for real human users is an open question. One obvious alternative is simply to ask for the same information again but in a rather different way. However, the set of possible actions in the learning experiment did not include this particular move.

Pietquin and Renals (2002) earlier trained a slot-filling system in which not all slots were required for a database query and generated a strategy which preferentially asked for slots whose values were more likely to be recognizable - a seemingly simple and effective policy which later researchers do not seem to have pursued.

Re-raising questions differently has also been proposed from the more theoretical strand of work using the Information State Update approach. Cooper and Larsson (2005) maintain a Qud-like stack of questions in the Information State which have been raised but not yet resolved. One possible use of this is so that later question repeats could be reformulated. Another is to allow interpretations of material which are not interpretable "alone" but require the earlier question as context. If the question remains suitably salient even though it is not the latest utterance then the material can be resolved. There is also the possibility of an accommodation mechanism in which material that requires a question to be salient

$$\begin{bmatrix} \text{FROM} \\ \text{DEPARTURE-TIME} \\ \text{TO} \\ \text{MODE OF TRAVEL} \end{bmatrix}$$

Figure 1: Travel Frame

1. [S] When do you want to leave?
2. [U] *I want to be in Gothenburg at 8.*

Figure 2: Unsolicited information

in order to be interpretable at all actually causes that question to become salient. Although demonstration systems have been built that are capable of illustrating these principles, it is far from clear that they are sufficiently robust to permit real dialogues to proceed smoothly and transparently to successful conclusions. Furthermore, discussions of these systems tend not to be explicit about how agents cost and select their moves.

3 Non-answers to questions

The key merit in the VoiceXML form interpretation algorithm is simply that it easily permits appropriate future dialogue behaviour upon encountering unsolicited relevant information, as in figure 2.

If "I want to be in Gothenburg at 8" can be understood as a possible answer to "How can I help you?" or even as an initiating exchange utterance on its own, then it ought also to be processable as a response to "When do you want to leave?". The user response is not an answer, not a direct one at least, but it can be used to fill in the value of a slot nonetheless. Then, the next question to be asked can be calculated as before on the basis of what else the system needs to know. The default VoiceXML strategy will end up repeating the very same question. The focus switching strategy (see above) will also repeat the very same question again only perhaps not just yet.

What is required for an agent to select more intelligently amongst the options it actually has available? In what circumstances would an agent move on to a different question. An agent rea-

sonably requests the same information (possibly via a re-phrasing) if the answer is either *essential* for progress; or if, more mundanely, it just remains the best bet for making progress. One advantage in concentrating on a typical task for a spoken dialogue system is that reasonable measures are easier to come by. If the immediate goal is to make a database query then an "essential" piece of knowledge is one without which no query can be made. A dialogue manager which insists on asking a particular question even though the underlying query system does not require it is clearly deficient. If the question is not essential, then the likelihood of making progress through a repeat needs to be *weighed*. Theoretical models stand in need of an explicit cost model and evaluation procedure in order to meet this demand.

Given the simplest slot-and-filler information state (Figure 1), one very simple addition to enable more intelligent selection is to just define a preference ordering over all the subsets of possible attributes e.g. $Sel : Pow(A) \rightarrow Z^{+}$. The selection algorithm then becomes: choose the attribute whose addition to those already supplied with values maximizes the value of Sel. Such a function is very easily implemented. Depending on the particular Sel function defined, a dialogue manager can now

1. repeat the original question
2. ask a different one
3. execute the database query straight away

If the system knows X and requests M, but receives N, then (X+N+M) might be considerably lower than (X+N+O) for some other O, so M should not be repeated but replaced by O. Indeed, X+N might be better than any extension of it, in which case no further questions should be asked: the intended database query can be made now. If we further add in a simple cost function which progressively penalizes repetitions of questions, then we essentially have the system of (Lewin, 2001). Another addition to the cost penalty might be the likelihood of receiving a recognizable answer, in the style of (Pietquin and Renals, 2002).

The cost and evaluation model can therefore be used to build in certain simple dialogue strategic principles. For example, the simple cost on repetitions allows repeats if the information sought is sufficiently important. Equally, there may be many sets of attributes which cannot be extended to a more preferred set. That is, there may be many different ways of achieving the goal of making *some* database query. Not all travel queries will require finding out the departure-time. Other general constraints might be imposed on the preference ordering. In general, more query constraints are better than fewer, because the set of satisfiers will likely be smaller and it will be easier to discuss and evaluate a small set of alternatives later. On the other hand, a query that is likely to return no answers (because the query is overly specified) might receive a very low value indeed.

Is this not precisely the sort of information that might be built into a reward function for a reinforcement learner? The answer of course is "yes" and "no". In principle, everything can be built in. In practice, not everything can be. Furthermore, what is built in in practice needs to be guided properly by theory. We have seen one instance of this already in which the learner learns to repeat a question later given that it prefers not to repeat it immediately but does actually need the answer to progress. Another highly plausible scenario is this: the travel options for Gothenburg can change over time and thus the best way to make progress in dialogues about those travel options can change over time too. As more travel options become available, it might become sensible to ask about the travel-mode earlier in the dialogue in order to reduce the size of the response from the database query. It is certainly feasible that a learning algorithm could learn automatically the changing relative costs of different travel queries; and these values could then function as an *input* to a dialogue management algorithm such as that sketched above. In general, the point is this: the development of theory which has practical ambitions needs to incorporate a cost model; furthermore, these developments can usefully *inform* further efforts at deploying machine learning. The point resembles that of Scheffler and Young: one need not attempt to model all possible dialogues but restrict the learning to parameters that one can effectively obtain data for.

The general analysis is not restricted to this partic-

ular scenario in which progress is evaluated by anticipated results of a database query. One might, for example, be able to execute a query after every new piece of information from the user and thereby have access to the actual travel possibilities given what the user has said so far. In this case, the next question might be calculated using Information Theoretical considerations: which question is the best next one to ask to identify the right travel option out of this set of possibilities? Again, values for such a parameter might be trainable from data without having to retrain the entire dialogue manager.

3.1 Stacks and Quds

Theoretical work in the Information State Update approach has often placed emphasis on storing questions (or issues) in stack like structures.

When the dialogue manager interprets the user's unsolicited offering and calculates his next move, need it note that a question was asked but not answered? Do we pop the question off a stack? The system knows of course which goals remain unsatisfied so the issue here is not whether to put the question again or not; just whether to note that there is an outstanding question that has not been answered. The distinction between public shared parts of dialogue state and private personal ones becomes important here. One intuition in the theoretical work is that the putting of a question is a public act that alters a shared linguistic environment; and this is independent of any mental states that might have led to its putting. Unfortunately the public or private status of an "outstanding" question is far less clear. Even in our simple extract (figure 2), the response might be a simple rejection of the question, in which case it is unclear if the original question remains outstanding or not. Alternatively, perhaps the user does intend to return to the original question at some point in the future and has placed the issue on a mental stack. Cooper and Larsson (2005) note how subtle a decision on this question might be.

I want to make two points about this issue. The first is that whatever our *best* interpretation of the user's offering, our next action should not simply be determined by that interpretation plus our own plans and goals. A rational cost calculation must include not only the risk of our best interpretation being incorrect but the possible cost of the next possible actions. How much will next actions differ if we think the user was rejecting our question or just postponing it. At best, we will only have a probability distribution over interpretations since the actual state of the user is of course unobservable. (Williams and Young, 2005) have recently been exploring the use of partially observable Markov decision processes in the settings of spoken dialogue systems.

The second, and strongly related, point is to emphasize that whether the dialogue "needs" to pop a stack or not is partly a matter of what strategic calculations about dialogue progress the dialogue manager *can* make. The dialogue manager, when interpreting the user's unsolicited offering, is of course also just about to make another contribution to the dialogue itself. If it is really only capable of asking questions at this point, then whatever the current state, the new most salient outstanding question will be the one it now chooses to make. Could the original question be required as context for a subsequent elliptical utterance? This is perhaps not impossible although it is unlikely given that the original question was not answered and has now been superseded by a new one. The original question is otiose as context if the question is just repeated of course.

The most important reason to record the original question on a stack is simply if the moves available to the dialogue manager include one that explicitly hands the initiative to the user. It is a common enough human strategy, if a non-answering response is made to a question, only to acknowledge the response. One might either signal, perhaps through intonation, that more input is expected or just wait and see what turns up next. This is not just a characteristic of chit-chat conversations, in which it doesn't particularly matter what happens next. It can be a calculated move in a strictly goal-driven interchange. In a chess game, one player might believe he understands his opponent's plan of attack, but play an inconsequential move to allow his partner the opportunity of making a more revealing move. Of course, such a move is a luxury item in the current state of spoken dialogue systems! Theoretically, however, the point is that "only acknowledging" is itself a particular dialogue move with consequences that need to be weighed in a model of move selection. Furthermore these possibilities are intimately tied to the addition of stack-like structures to the in-

1. [S] Do you want the train or the plane?
2. [U] *What time do they arrive?*

Figure 3: Dependent questions

1. [S] Do you want the train or the plane?
2. [U] *What time do they arrive?*
3. [S] The train arrives at 3 and the plane at 4
4. [U] *The train please*

Figure 4: Long distance short answers

formation state. If a dialogue manager is to be *capable* of just acknowledging a user assertion, then it ought also to be sensitive to the cost of doing that versus something else. Clearly, practical dialogue managers with this sort of capability have yet to be built. Theoretical accounts that include acknowledgments often appear to include them simply as part of a "protocol of interaction".

4 Questions in answer to questions

Stacks and Quds are popular also for analysis of dialogues that include nested questioning sequences (figure 4). Classically, these cases are ones where the answer to the first question somehow depends on the answer to the second. If it is analysed as being dependent, then the first question may be put on a stack; or perhaps a nested conversational game is begun, or possibly an outstanding obligation to answer it is recorded (Ginzburg, 1996; Lewin, 2000; Matheson et al., 2000).

Shifting the focus of our attention onto next move selection again urges rather the importance of weighing the possible next moves. What should a dialogue manager do when it receives a question in a response to question? Clearly, if the goal motivating the original question is still of primary importance, then it might be best simply to re-ask it (possibly with a re-phrasing). Alternatively, perhaps the question can be re-asked whilst also answering the new question. Finally, perhaps it is best simply to move onto another topic altogether. As we saw earlier with acknowledgments, it would be a mistake to suppose that the new question must simply be answered because of a protocol of interaction. In fact, the tactic of *only* answering the new question is just one further specific move with its own specific advantages and disadvantages that need to be weighed.

Does the dialogue manager need to record the original unanswered question? Again, the first point to make is that the original goal is still outstanding so the issue is one of noting an unanswered question, rather than whether to re-ask it. In the current case, it is clearer that the question might function as a context for a possible later ellipsis resolution; and this is arguably what happens in figure 4. Nevertheless, the matter is again intimately tied to the dialogue manager's choice of move, namely *only answering* the interjected question. The question of what to include in the dialogue state is not independent of what choices of action the dialogue manager can make. Practically speaking of course, it is worth remembering that correct ellipsis resolution is in any case not a simple operation. It is also increasingly hard the longer the distance grows between context and ellipsis and that, in many situations, there may often be other ways to achieve the right interpretation. Perhaps in most travel scenarios, "the train" only ever means "I'd like to travel by train".

Ginzburg, in his careful theoretical analysis (Ginzburg, 2007), also notes that dialogue stack structures do not arise as the results of seemingly arbitrary rules of dialogue syntax but from underlying dependencies amongst issues. That is, the answer to "train or plane" may depend upon arrival times. He goes further and states that "Since B can be assumed to attempt to provide information about $q1$ were she able to do so, the fact that she has not done so must mean that B cannot do so without (attempting to) resolve $q2$". It seems to me that this is just not true. Certainly in educational contexts, a teacher may respond to a question with another because this is just the best way to further the underlying educational objective even though a direct answer could have been given. This is not uncooperative behaviour. If a question is not answered, it does not follow that it could not have been answered, merely that the agent's choice of his next best move to further his own interests was not to answer it.

5 Cooperation and cost

A rather rich example of progress in a non-question answering dialogue is illustrated in Figure 5, a genuine extract from a travel agency dialogue conducted

1. [S] Välkommen till resebokningstjänsten. Hur jan jag hjälpa dig?
   [Welcome to the travel bureau, How can I help you?]
2. [U] Jag vill ha en resa till Göteborg från Stockholm.
   [I want to travel from Stockholm to Gothenburg.]
3. [S] Vilket datum vill du resa?
   [What day will you travel on?]
4. [U] Jag vill vara framme i Göteborg den elfte i åttonde.
   [I want to be in Gothenburg on August 11th.]
5. [S] Vill du resa med tåg eller flyg?
   [Will you travel by plane or train?]
6. [U] Vilken tid kommer dom fram?.
   [What time do they get there?]
7. [S] Vilken tid vill du resa ungefär?
   [What time will you leave?]
8. [U] Jag vill vara framme i Göteborg före klockan åtta.
   [I want to be in Gothenburg for eight o'clock]
9. [S] Det finns ett flyg som avgår klockan sju noll noll och ankommer klockan åtta noll noll. Det är det tidigaste flyget.
   [There is a flight that leaves at seven zero zero and arrives at eight zero zero. That is the earliest flight.]

Figure 5: Progress without Answers

between a human and a human (or "wizard") simulating a computer system. The wizard used a script to determine the next system utterance although it did not necessarily cover all conceivable circumstances.

The dialogue complexity increases rapidly after utterance 4. It is an interesting exercise to try to discern which questions remain open for discussion at which points and which cues humans might use to help decide. Could the behaviour be reproduced by a wholly automatic computer system in conversation with a human? What structures would it need to do it? How much inference would it need to employ? The answer is simple: it only requires the simple resources we have already sketched. Let us suppose the wizard is actually issuing a database query after each user utterance and choosing what to say next based partly on the size of the results of that query (the "solution set"). Starting at 5, the wizard determines that an answer to the question of travel mode will most likely reduce the size of the solution set the most. The user does not answer this question. Unfortunately, his offering, 6, also does not reduce the solution set size at all; although had he asked "How much do they cost?", the story might be different as there may be many fewer costs than travel options. The number of arrival times in the solution set may similarly be too high. What should the wizard do next? Repeating the travel mode question is a possibility; and is still presumably optimal with respect to solution set size; but there is a penalty for repeats. So, in this case, the next best question, 7, is decided upon. The user now offers 8. Response 8 also does not answer the previous question. But 8 does in fact reduce the size of the solution set sufficiently and this phase of the dialogue can successfully close through 9.

It is noteworthy that none of the three questions, 5,6 and 7 was actually answered. Yet the dialogue has succeeded. Furthermore, the suggested flight in 9 is not, I think, even an answer to questions 5, 6, and 7. Indeed, even if the user were to follow up with "I'll take it", it is far from clear that that is a response to 5, 6, and 7, as well as 8.

Why could motivate a user to follow up question 5 with his own question but then question 7 with a non-answering statement? It appears that the user's motivating goal was probably all along the desire to be in Gothenburg by eight o'clock on August 11th. He was happy to play along with the system's line of questioning so long as progress was also being made towards this goal. Question 5 put this in jeopardy. An answer to this question might inadvertently rule out the possibility of being in Gothenburg at the desired time. By asking for arrival times, the user planned to pick one before eight o'clock. This tactic failed. However, the system's next question also could not further his objective and so he decided not to play along with the system's strategy anymore. He decided to override the system's question entirely and state the hard constraint he had in mind. That is one possible interpretation at any rate.

Does the dialogue instantiate *uncooperative* be-

haviour? Certainly questions were left unanswered; but each agent did at all times attempt to advance his own goals and, as the two sets of goals were indeed related, a mutually agreeable path forwards could be found. Co-operation is simply something that can exist at different levels of activity and at different times. One can answer a question in a way that does not advance the task; just as one can advance the task by not answering a given question. Cooperation is not a complete package that one just buys into or not in any given conversation. Besides, a robust system needs to be able to cope with a mildly uncooperative human it encounters.

6 Conclusion

Explicit cost models form an essential part of a complete account of dialogue management in the Information State Update model. The information required in order to make dialogue moves is just as important as that required to interpret them. Furthermore building in certain interpretative capabilities (such as stacked questions for ellipsis resolution) actually depends on the set of moves that a dialogue manager can make (such as "only acknowledging"). Finally, the addition of cost models and the design of strategy is not exhausted by the possibility of machine learning scenarios. Indeed, good theory can help direct the machine learning towards acquiring valuable parameters more effectively.

Acknowledgments

Thanks to Telia Research AB, Vitsandsgatan 9, S-123 86 Farsta, Sweden for permission to use their corpus of travel dialogues.

References

J. Chu-Carroll and J. Nickerson. 2000. Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In *Proc. 1st NAACL*, pages 202–209.

R. Cooper and S. Larsson. 2005. Accommodation and reaccommodation in dialogue. In R. Bäuerle, U. Reyle, and T.E. Zimmerman, editors, *Presuppositions and Discourse*. Elsevier.

Matthew Frampton and Oliver Lemon. 2006. Learning more effective dialogue strategies using limited dialogue move features. In *Proc. ACL '06*, pages 185–192.

J. Ginzburg. 1996. Interrogatives: Questions, facts, and dialogue. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 385–422. Blackwell Publishers.

J. Ginzburg. 2007. *Semantics and Interaction in Dialogue*. Studies in Computational Linguistics. CSLI Publications.

J. Henderson, O. Lemon, and K. Georgila. 2005. Hybrid reinforcement/supervised learning for dialogue policies from communicator data. In *IJCAI workshop: Knowledge & Reasoning in Practical Dialogue Systems*.

Oliver Lemon, Kallirroi Georgila, and James Henderson. 2006. Evaluating effectiveness and portability of reinforcement learned dialogue strategies with real users. In *IEEE/ACL Spoken Language Technology*.

I. Lewin. 2000. A formal model of conversational game theory. In *4th Workshop on the Semantics & Pragmatics of Dialogue: Gotalog 2000*, pages 115–122.

I. Lewin. 2001. Limited enquiry negotiation dialogues. In *Eurospeech 2001: Proc. 7th European Conference on speech communication and technology*, pages 2333–2336.

Diane J. Litman and Shimei Pan. 2002. Designing & evaluating an adaptive spoken dialogue system. *User Modeling & User-Adapted Interaction*, 12(2-3):111–137.

C. Matheson, M. Poesio, and D. Traum. 2000. Modelling grounding and discourse obligations using update rules. In *Proceedings of NAACL 2000*.

Matt Oshry, RJ Auburn, Paolo Baggia, Michael Bodell, David Burke, Daniel C. Burnett, Emily Candell, Jerry Carter, Scott McGlashan, Alex Lee, Brad Porter, and Ken Rehor. 2006. Voice extensible markup language (voicexml) 2.1.

Olivier Pietquin and Steve Renals. 2002. Asr system modeling for automatic evaluation and optimization of dialogue systems. In *Proceedings of ICASSP*, volume I, pages 45–48, Orlando, (USA, FL).

Konrad Scheffler and Steve Young. 2002. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proc. 2nd Int. Conf. on Human Language Technology Research*, pages 12–19. Morgan Kaufmann Publishers Inc.

J. D. Williams and S. Young. 2005. Scaling up pomdps for dialog management: The "summary pomdp" method. *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 177–182.

Andrzej Wiśniewski

Section of Logic and Cognitive Science
Institute of Psychology
Adam Mickiewicz University in Poznań, Poland
Andrzej.Wisniewski@amu.edu.pl

QUESTIONS, INFERENCES, AND DIALOGUES

1. Internal Question Processing

Logical theories of questions supply formalisms for questions as well as characteristics of the question-answer relation.[1] As long as question asking and question answering are concerned, they usually adopt a simple dyadic perspective. It is assumed that there are two parties: a questioner and an answerer. The former asks a question, whereas the role of the latter is to provide an answer to the question. Even eliciting information from Nature is modeled that way.[2] When cooperative questioning is analyzed, an agent can play both roles, depending on the stage. This dyadic perspective, however, seems to obscure some important phenomenon, which may be called ***internal question processing***.

What we mean here by internal question processing (hereafter: **IQP**) is not tantamount to question answering. When a question is internally processed, the immediate outcome need not be an answer to this question: an 'inference' performed on a question can lead to another question, which may be 'send' by a cognitive agent either to itself or to a certain external source of information and then answered, but can also be processed further in an analogous way. Usually, this results in a problem decomposition: a (difficult) problem represented by a certain question is decomposed into sub-problems represented by other questions. However, the decomposition is dynamic and comes in stages: the consecutive questions (which constitute the sub-goals of the next stage) depend on how the previous requests for information have been fulfilled. In other cases a (difficult) problem represented by a certain question is restated by formulating a new question. The relevant transformations of questions usually facilitate question answering and problem-solving. But there are cases in which they result in a plausible answer/solution to a question/problem.

In brief, our main objective will be to present some logical tools which are useful in a formal modeling of **IQP**.

2. Erotetic Inferences

In order to provide a formal account of **IQP** we need a logic which analyzes inferences performed on questions and proposes criteria of their validity. At first sight this claim may seem a contradiction, since questions are neither true nor false. But a moment's reflection shows that there are inferential thought processes which result in questions. They are called ***erotetic inferences*** (from Greek 'erotema', which means 'question').

Sometimes we pass from proposition(s) to a question, as in:

(1) *Andrew always comes in time, but now he is late. So what has happened to him?*

We also pass from a question to a question on the basis of some proposition(s), *e.g.*:

(2) *Where did Andrew leave for? If Andrew took his famous umbrella, then he left for London; otherwise he left for Paris or Rome. So did Andrew take his famous umbrella?*

Moreover, it happens that we pass from a question directly to a question, as in:

(3) *Is 112657853 a prime? So is there a natural number divisor of 112657853 different from it and from 1?*

In the second and third cases inferences are performed on questions: they play the roles of 'premises' and 'conclusions'.

Inferential Erotetic Logic (**IEL** for short) puts erotetic inferences in the centre of

[1] For an overview see: Harrah, D., 'The Logic of Questions', in: D. Gabbay, T. Guenthner (eds.) *Handbook of Philosophical Logic, Second Edition*, Volume 8, Kluwer, Dordrecht/ Boston/ London 2002, pp. 1-60.

[2] *Cf.* Hintikka, J., *Inquiry as Inquiry: A Logic of Scientific Discovery*, Kluwer, Dordrecht/ Boston/ London 1999.

its interest.[3] **IEL** gives an account of the phenomenon of question raising and defines validity of erotetic inferences.

As for question raising, we have two different types of cases here, corresponding to two types of erotetic inferences. A question can arise out of a set of propositions, and a question can arise from a question on the basis of a (possibly empty) set of propositions. The relevant concepts of question raising are explicated in **IEL** by defining the semantic concepts of ***evocation*** of a question by a set of propositions, and ***erotetic implication*** of a question by a question and a set of propositions. A semantic approach is then mirrored by a syntactic one, and question-evoking and question-implying rules are formulated. ***Validity*** of erotetic inferences is defined in terms of evocation and erotetic implication, respectively. Erotetic implication, as characterized in **IEL**, has a 'teleological' feature: an implied question Q^* is not only semantically grounded in the implying question Q, but Q^* is also cognitively useful with respect to Q in an 'open-mined' way: *i.e.* each direct answer to Q^* potentially contributes to finding, at least partial, answer to Q. Let us stress that the latter condition is explicated in semantic terms.

We will concentrate upon erotetic inferences which have questions as premises and conclusions, and thus on erotetic implication. This relation will be defined in terms of the so-called Minimal Erotetic Semantics.

3. Distributed IQP and E-Scenarios

One can distinguish two types of **IQP**: ***ultimate*** and ***distributed***.

As long as ultimate **IQP** is concerned, no information requests are sent and the processing itself may lead to a plausible answer to a question. In the case of distributed **IQP** requests for additional information are sent, and questions are transformed into further questions depending on how previous information requests have been fulfilled. These requests for information may be sent by a cognitive agent to itself (for instance, in order to activate his/her memory), or to a certain external source of stored information, or to other cognitive agent (*e.g.* in an information-seeking dialogue).

The concept of ***erotetic search scenario*** (e-scenario for short) can be useful in the formal modeling of **IQP**. [4]

An e-scenario is an abstract structure defined by means of tools taken from **IEL**. However, an e-scenario function is to show how a principal question may be answered by asking and answering auxiliary questions. An e-scenario has a tree-like structure with the principal question as the root and possible answers to this question as leaves. Other questions enter e-scenarios on the condition they are erotetically implied (in the sense of **IEL**). Moreover, an auxiliary question either: (a) has another question as the immediate successor, or (b) all the direct answers to the auxiliary question follow the question as its immediate successors. In the latter case an auxiliary question is a ***query*** and the immediate successors represent the possible ways in which the relevant request for information can be satisfied. The structure of an e-scenario shows what kind of further information requests (if any) are to be satisfied in order to arrive at an answer to the principal question.

Distributed **IQP** can be modeled in terms of e-scenarios in various ways. One of the possible lines of thought is the following. We attribute to a cognitive agent an initial e-scenario for his/her principal question just processed. The topmost query of this e-scenario determines the first request for information to be sent. Now, when the query is answered in a given way, the e-scenario ***contracts***: consecutive queries which would follow the alternative answers to the query become inessential, and one arrives at a new e-scenario (again, for the principal question) with a new 'topmost' query, which is processed analogously. But suppose that one arrives at a query such that no answer to it is available by existing means. So, a ***revision*** of the current e-scenario is needed. One possible move is a revision by ***embedding***: an e-scenario for the

[3] Cf. Wiśniewski, A., *The Posing of Questions: Logical Foundations of Erotetic Inferences*, Kluwer, Dordrecht/ Boston/ London 1995, or: Wiśniewski, A., 'The logic of questions as a theory of erotetic arguments', *Synthese* 10**9**, No. 2, 1996, pp.1-25; Wiśniewski, A., 'Questions and inferences', *Logique et Analyse* 173-175, 2001, pp. 5-43.

[4] *Cf.* Wiśniewski, A., 'Erotetic search scenarios', *Synthese* 134, No. 3, 2003, pp. 389-427; see also: Wiśniewski, A., 'Erotetic search scenarios, problem-solving, and deduction' *Logique et Analyse* 185-188, 2004, pp. 139-166.

troublemaking query is embedded into the e-scenario just considered. Another possible move is a revision by ***conditionalisation***: an answer to the query is added (with an appropriate comment) to the initial premises and the current e-scenario contracts accordingly. There are also other moves possible. Note that it is the initial e-scenario that is being transformed. As a consequence, the following desirable property is retained: each path of an intermediate scenario leads to an answer to the principal question. The process as a whole is goal-directed, comes in stages, and the sub-goals are processed/ created in a dynamic way.

The concept of e-scenario will be introduced, some operations of e-scenarios will be characterized, and the issue of applicability of the concepts of erotetic implication and e-scenario in the analysis of dialogues will be discussed.

Commitments, Beliefs and Intentions in Dialogue

Nicholas Asher
IRIT
Université Paul Sabatier, Toulouse
asher@irit.fr

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

Abstract

We define grounding in terms of shared public commitments, and link public commitments to other, private, attitudes within a decidable dynamic logic for computing implicatures and predicting an agent's next dialogue move.

1 Introduction

A theory of dialogue should link discourse interpretation to general principles of rationality and cooperativity (Grice, 1975). The so-called 'mentalist approach' treats dialogue as a function of the agents' attitudes, usually formalised with BDI (belief, desire, intention) logics (e.g., Grosz and Sidner (1990)). Grounding a proposition p—by which we mean that all dialogue agents mutually agree that p is true—occurs when the BDI logic implies that p is mutually believed.

However, there are compelling reasons to reject the mentalist approach to dialogue modelling. Gaudou et al. (2006) use (1) to argue for a distinction between grounding and mutual belief.

(1) a. A to B (C out of earshot): C is stupid.
b. B to A (C out of earshot): I agree.
c. A to B (C in earshot): C is smart.

(1a) is grounded for A and B. If B now utters *That's right*, then (1c) should be grounded for A and B too. So if grounding is a function of mutual belief, then A and B would hold contradictory beliefs, making them irrational. But A is not irrational; he is disingenuous. Gaudou et al. (2006) conclude that grounding is a function of *shared public commitments*, following Hamblin (1987). But the link to other attitudes is also essential: B should detect that A is lying—i.e., that he can't believe everything that he has publicly committed to.

Dialogue (1) contrasts with dialogue (2), where A 'drops' a commitment to (2a) in favour of (2b), making (2b) grounded:

(2) a. A: It's raining.
b. B: No it's not.
c. A: Oh, you're right.

A theory of dialogue should distinguish between A's illocutionary act in (1c) vs. (2c), even though in both cases A asserts the negation of his prior assertion.

In this paper, we propose a framework for dialogue analysis that synthesises Hamblin's commitment-based approach with the mentalist approach. We think both perspectives on dialogue are needed. In Lascarides and Asher (2008), we argue that the commitment-based framework captures facts about grounding, making explicit the distinction between what is said and private attitudes. But the BDI view is essential for strategic reasoning about dialogue moves. We draw on the strengths of both approaches while avoiding some of their weaknesses. For instance, we avoid the uncomputable models of discourse that stem from default reasoning in first-order BDI logics.

Our starting point is SDRT (Asher and Lascarides, 2003). In Section 2 we modify its representation of dialogue content so that it tracks the public commitments of each dialogue agent. In Section 3 we reconstruct its separate, but related, cognitive logic (CL) to include the attitude of public commitment and axioms that relate it to other, private, attitudes. CL will be a dynamic logic of public announcement, extended with default axioms of rationality and cooperativity. The result will capture the sort of prac-

Turn	A's SDRS	B's SDRS
1	$\pi_1 : K_{\pi_1}$	$\emptyset$
2	$\pi_1 : K_{\pi_1}$	$\pi_{2B} : \textit{Explanation}(\pi_1, \pi_2)$

Table 1: The logical form of dialogue (3).

tical reasoning that goes on in conversation, when agents adjust their beliefs, preferences and intentions in light of what's said so far. This refines the approach to dialogue using dialogue games (e.g., Amgoud (2003)) because the utilities for each possible dialogue move need not be 'pre-defined' or quantified. Rather, CL will exploit the dynamics in the logic to infer qualitative statements about the *relative* utility of different moves. Furthermore, by approximating game-theoretic principles within the logic, we also deepen the theory by *deriving* some of the cognitive axioms of rationality and cooperativity from them: for instance, a general axiom of Cooperativty (that B normally intends what A intends) will be validated this way. Our approach can also be viewed as extending the Grounding Acts Model (Traum, 1994), providing its update rules with a logical rationale for constraining the update effects on content vs. cognitive states.

2 Dialogue Content

Lascarides and Asher (2008) argue that relational speech acts or *rhetorical relations* (e.g., *Narration*, *Explanation*) are a crucial ingredient in a model of grounding. One of the main motivations is implicit grounding: representing the illocutionary contribution of an agent's utterance via rhetorical relations reflects his commitments to another agent's commitments, even when this is linguistically implicit. For example, B's utterance (3b) commits him to (3a) because the relational speech act $\textit{Explanation}(3a, 3b)$ that he has performed entails (3a):

(3) a. A: Max fell.
b. B: John pushed him.

Accordingly, the commitments of an individual agent are expressed as a Segmented Discourse Representation (SDRS, Asher and Lascarides (2003)): this is a hierarchically structured set of labelled contents, as shown in each cell of Table 1—the logical form for dialogue (3). For simplicity, we have omitted the representations of the clauses (3a) and (3b) (labelled π_1 and π_2 respectively), and we often gloss the content labelled by π as K_π, and mark the root label of the speaker i's SDRS for turn j as π_{ji}.

The logical form of dialogue is the logical form of each of its turns (where a turn boundary occurs whenever the speaker changes). The logical form of each turn is a set of SDRSs, one for each dialogue participant. Each SDRS represents *all* the content that the relevant agent is currently publicly committed to, from the beginning of the dialogue up to the end of that turn (see Lascarides and Asher (2008) for motivation). And each agent constructs the SDRSs for all other agents, as well as his own—e.g., A and B both build Table 1 for dialogue (3).

The logical form of dialogue (2) is Table 2. Recognising that B's utterance π_2 attaches to π_1 with *Correction* is based on default axioms in SDRT's *glue logic*—i.e., the logic for constructing logical form (Asher and Lascarides, 2003). The content of (2c) (labelled π_3) supports a glue-logic inference that π_3 acknowledges π_2. This resolves π_3's underspecified content to entail K_{π_2}, and so $\textit{Correction}(\pi_1, \pi_3)$ is also inferred, as shown. In contrast, the fact that (1c) is designed to be overheard by C while (1ab) is not forces a glue-logic inference that they are not rhetorically linked at all; see the logical form in Table 3.

The dynamic semantics for Dialogue SDRSs (DSDRSs) is defined in terms of SDRSs: a DSDRS consists of an SDRS for each participant at each turn, and accordingly the semantics of a dialogue turn is the product of the dynamic semantics for each constituent SDRS. Lascarides and Asher (2008) define grounding at a given turn as the content that's entailed by each SDRS for that turn. Given that each turn represents *all* an agent's 'current' public commitments, the interpretation of a dialogue overall is that of its last turn. Table 2 receives a consistent interpretation, but Table 3 is inconsistent because A's final SDRS is inconsistent. The DSDRS of (3) makes (3a) grounded and that for (2) makes (2b) grounded. The DSDRS of (1) makes (1a) grounded, and should B acknowledge (1c), then anything is grounded.

3 Cognitive Modelling

With this background concerning dialogue content in place, we turn to the interaction of commitments with other attitudes. SDRT's cognitive logic (CL)

Turn	A's SDRS	B's SDRS
1	$\pi_1 : K_{\pi_1}$	$\emptyset$
2	$\pi_1 : K_{\pi_1}$	$\pi_{2B} : \textit{Correction}(\pi_1, \pi_2)$
3	$\pi_{3A} : \textit{Correction}(\pi_1, \pi_3) \wedge \textit{Acknowledgement}(\pi_2, \pi_3)$	$\pi_{2B} : \textit{Correction}(\pi_1, \pi_2)$

Table 2: The logical form of dialogue (2).

Turn	A's SDRS	B's SDRS
1	$\pi_1 : K_{\pi_1}$	$\emptyset$
2	$\pi_1 : K_{\pi_1}$	$\pi_{2B} : \textit{Acknowledgement}(\pi_1, \pi_2)$
3	$\pi_{3A} : K_{\pi_1} \wedge K_{\pi_3}$	$\pi_{2B} : \textit{Acknowledgement}(\pi_1, \pi_2)$

Table 3: The logical form of (1).

supports reasoning about agents' cognitive states in virtue of what they say. Since it contributes directly to constructing the logical form of dialogue, its complexity must be decidable: Asher and Lascarides (2003, p78) argue that this is necessary to explain why, as Grice (1975) claims, people by and large agree on what was said (if not on whether it's true). CL must support default reasoning and hence consistency tests, since agents never have complete information about the dialogue context. And so SDRT makes its CL decidable by denying it access to a dialogue's full, dynamic interpretation—for instance, existentially-quantified SDRS-formulae lose their structure when transferred into CL, thereby losing the relationship between, say, the SDRS-formulae $\neg\exists x\neg\phi$ and $\forall x\phi$.

SDRT's CL from Asher and Lascarides (2003) is deficient in at least two ways. First, it does not support the logical forms from Section 2; CL should include public commitment and its links to other attitudes. Secondly, CL is static, thereby failing to show how attitudes change during dialogue. To overcome these deficiencies we exploit a dynamic logic of public announcement (Baltag et al., 1999). We extend it to support *default* reasoning from public announcements, including (default) inferences about cognitive states. A model $\mathcal{M}$ of the logic consists of a set of worlds $W^{\mathcal{M}}$ and a valuation function $V^{\mathcal{M}}$ for interpreting the non-logical constants at $w \in W^{\mathcal{M}}$. We write $[\![\phi]\!]^{\mathcal{M}} =_{def} \{w \in W^{\mathcal{M}} : \mathcal{M}, w \models \phi\}$. Public announcements are dynamic in that they change the input model into a different output one: any worlds from the input model which fail to satisfy the monotonic consequences of the announcement are eliminated from the output model; likewise for *ceteris paribus* announcements, any worlds that fail to satisfy the nonmonotonic consequences of the announcement are eliminated. More formally, monotonic consequences of an announcement are expressed by the formula $[!\phi]\psi$, where $[!\phi]$ is a modal operator (in words, ψ follows from announcing ϕ). Nonmonotonic consequences are expressed as $[!\phi]^{cp}\psi$, which in turn is defined via a modal connective: $\phi > \psi$ means that *If ϕ then normally ψ*. The model $\mathcal{M}$ therefore also includes a function $*$ from worlds and propositions to propositions, which defines normality and is used to interpret $\phi > \psi$:

$$\mathcal{M}, w \models \phi > \psi \text{ iff } *^{\mathcal{M}}(w, [\![\phi]\!]^{\mathcal{M}}) \subseteq [\![\psi]\!]^{\mathcal{M}},$$

In words, ψ is true in all worlds where, according to w, ϕ is normal. The above description of how announcements transform input models is then formalised in Figure 1.

$$\begin{array}{rl}
\mathcal{M}, w \models [!\phi]\psi \text{ iff} & \mathcal{M}^{\phi}, w \models \psi \\
\mathcal{M}, w \models [!\phi]^{cp}\psi \text{ iff} & \mathcal{M}^{cp(\phi)}, w \models \psi \\
\text{where} & \\
\mathcal{M}^{\phi} = & \langle W^{\phi}, *^{\mathcal{M}}|W^{\phi}, V|W^{\phi}\rangle \text{ where} \\
W^{\phi} = & [\![\phi]\!]^{\mathcal{M}} \\
\mathcal{M}^{cp(\phi)} = & \langle W^{cp(\phi)}, *^{\mathcal{M}}|W^{cp(\phi)}, V|W^{cp(\phi)}\rangle \text{ where} \\
W^{cp(\phi)} = & \{w' \in W^{\mathcal{M}} : \\
& \quad \text{Th}(\mathcal{M}), \phi \mathrel{|\!\sim} \psi \rightarrow \mathcal{M}^{\phi}, w' \models \psi\}
\end{array}$$

Figure 1: Model transitions for announcements

To ensure that CL reflects the commitments in DS-DRSs, we assume that agents announce to the dialogue participants certain commitments to SDRS-formulae. Actually, given the way we have set things

up, each turn commits a speaker to commitments from earlier turns, unless he disavows one of those commitments. $\mathcal{P}_{a,D}\psi$ means that a publicly commits to group D to ψ. Thus a speaker a uttering K_π to D will result in CL-based reasoning with the modality $[!\mathcal{P}_{a,D}\phi_\pi]^{cp}$, where ϕ_π is the shallow representation of K_π (i.e., without existentials). We make the modality $\mathcal{P}_{a,D}$ K45 (one commits to all the consequences of one's commitments, and one has total introspection on commitments, or lack of them), and we also add axioms Ax1 (a commitment to D is a commitment to all its subgroups), and Ax2 (there is a group commitment by x and y to D iff x and y both make that commitment to D):

K: $\mathcal{P}_{a,D}(\phi \rightarrow \psi) \rightarrow (\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D}\psi)$
4: $\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D}\mathcal{P}_{a,D}\phi$
5: $\neg\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D}\neg\mathcal{P}_{a,D}\phi$
Ax1: For any $D' \subseteq D$, $\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D'}\phi$
Ax2: $\mathcal{P}_{\{x,y\},D}\phi \leftrightarrow (\mathcal{P}_{x,D}\phi \wedge \mathcal{P}_{y,D}\phi)$

So the models $\mathcal{M}$ have suitably constrained accessibility relations $R^{\mathcal{P}_{a,D}} \subseteq W \times W$ for all a and D.

Since commitment lacks axiom D, $\mathcal{P}_{a,D}(p \wedge \neg p)$ is satisfiable, reflecting A's public commitments in (1). This contrasts with the belief modality $\mathcal{B}_a\phi$, which is KD45 (with a transitive, euclidean and *serial* accessibility relation $R^{\mathcal{B}_a}$ in the model).

Agent a announcing something to group D will bring about in CL a transition on models: the input model will be updated by adding to a's commitments to D. Changing a model by adding ϕ to a's commitments is defined in equation (4): this stipulates that one adds ϕ to the accessibility relation $R^{\mathcal{P}_{a,D}}_{\mathcal{M}}$, so long as doing so is consistent. Equation (5) defines a similar model transition for beliefs; we'll use this shortly to represent Sincerity.

(4) $\mathcal{M} \mapsto \mathcal{M}_{\phi,a,D}: R^{\mathcal{P}_{a,D}}_{\mathcal{M}_{\phi,a,D}} = (?\top; R^{\mathcal{P}_{a,D}}_{\mathcal{M}}; ?\phi)$

(5) $\mathcal{M} \mapsto \mathcal{M}_{\flat_a\phi}: R^{\mathcal{B}_a}_{\mathcal{M}_{\flat_a\phi}} = (?\top; R^{\mathcal{B}_a}_{\mathcal{M}}; ?\phi)$

We can now interpret announcements about commitments. In words, should an agent a say ϕ to D, then the model is udpated so that all non-monotonic consequences of a's commitment to ϕ are satisfied (so long as this update is consistent):

- `Announcements of Commitment`:
  $\mathcal{M}, w \models [!\mathcal{P}_{a,D}\phi]^{cp}\psi$ iff $\mathcal{M}^{cp(\phi)}_{\phi,a,D}, w \models \psi$

In fact, we assume that should a say K_π to D, then in CL the *ceteris paribus* consequences of this announcement include a's commitment to all glue-logic inferences χ about the illocutionary effects of K_π (as represented via rhetorical relations in the DSDRSs): i.e., $[!\mathcal{P}_{a,D}\phi_\pi]^{cp}\mathcal{P}_{a,D}\chi$. This yields $[!\mathcal{P}_{B,\{A,B\}}\phi_{\pi_2}]^{cp}\mathcal{P}_{A,\{A,B\}}$*Explanation*$(\pi_1, \pi_2)$ in CL from dialogue (3), for instance. Thus the outcome in CL is a model that satisfies $\mathcal{P}_{B,\{A,B\}}$*Explanation*(π_1, π_2), and so long as enough of the semantics of *Explanation* is transferred into CL, this entails (by axiom K) $\mathcal{P}_{B,\{A,B\}}\phi_{\pi_1}$, where ϕ_{π_1} is the shallow representation (3a). A's announcement (3a) ensures the CL model also satisfies $\mathcal{P}_{A,\{A,B\}}\phi_{\pi_1}$. So the CL model reflects what's grounded according to the DSDRS. Table 2, the representation of dialogue (2), yields a CL model that satisfies $\mathcal{P}_{\{A,B\},\{A,B\}}\phi_{\pi_2}$ and $\mathcal{P}_{\{A,B\},\{A,B\}}\neg\phi_{\pi_1}$, where ϕ_{π_1} and ϕ_{π_2} represent (2a) and (2b) respectively. And Table 1 yields a CL model where $\mathcal{P}_{A,\{A,B\}}(p \wedge \neg p)$, p being the (shallow) CL representation of (1a).

An agent's beliefs must be updated at least defeasibly on discovering his commitments. The following `Sincerity` axiom ensures this, by default:

- `Sincerity`: $\mathcal{P}_{a,D}\phi > \flat_a\phi$

We have stated `Sincerity` dynamically via the action operator $\flat_a$; this is the action of updating beliefs and has the following semantics:

- `Belief Update`:
  $\mathcal{M}, w \models \flat_a\phi$ iff $\mathcal{M}_{\flat_a\phi}, w \models \mathcal{B}_a\phi$

`Sincerity` is a default because of examples like (1). As we saw earlier, `Announcements of Commitment` yields $\mathcal{P}_{A,\{A,B\}}(p \wedge \neg p)$. This satisfies the antecedent to `Sincerity`, but $\mathcal{B}_A(p \wedge \neg p)$ is not inferred because it's inconsistent. $\mathcal{P}_{A,\{A,B\}}p$ and $\mathcal{P}_{A,\{A,B\}}\neg p$ are also true (by axiom K); they both satisfy the antecedent of `Sincerity`, but their consequences $\mathcal{B}_A p$ and $\mathcal{B}_A\neg p$ are mutually inconsistent, and so neither is inferred. Thus B detects from A's inconsistent current commitments that he's lying, and without further information B does not know what A believes: p, $\neg p$ or neither one. C, on the other hand, who knows only $\mathcal{P}_{A,\{A,B,C\}}\neg p$, uses `Sincerity` to infer $\mathcal{B}_A\neg p$.

As is standard, mutual belief (*MB*$_{x,y}\phi$) is defined in terms of belief using a fixed point equation:

(6) *MB*$_{x,y}\phi \leftrightarrow (\mathcal{B}_x(\phi \wedge$ *MB*$_{x,y}\phi) \wedge \mathcal{B}_y(\phi \wedge$ *MB*$_{x,y}\phi))$

This definition means $MB_{x,y}\phi$ entails an ω sequence of nested belief statements: $\mathcal{B}_x\phi, \mathcal{B}_y\mathcal{B}_x\phi, \ldots$ and $\mathcal{B}_y\phi, \mathcal{B}_x\mathcal{B}_y\phi, \ldots$. We will denote a formula that starts with $\mathcal{B}_x$, and alternates with $\mathcal{B}_y$ to a nesting of depth n as $\mathcal{B}^n_{(x,y)}\phi$; similarly for $\mathcal{B}^n_{(y,x)}\phi$. Then one can prove the following scheme is sound.

- `Induction Scheme`:
  Assume $\Gamma \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}_x\phi) \wedge \mathcal{B}_x(\phi \wedge \mathcal{B}_y\phi)$
  And for any n, $\dfrac{\Gamma \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}^n_{(x,y)}\phi) \wedge \mathcal{B}_x(\phi \wedge \mathcal{B}^n_{(y,x)}\phi)}{\Gamma \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}^{n+1}_{(x,y)}\phi) \wedge \mathcal{B}_x(\phi \wedge \mathcal{B}^{n+1}_{(y,x)}\phi)}$
  Then: $\Gamma \mathrel{|\!\sim} MB_{x,y}\phi$

These axioms ensure that, as in the BDI account, grounding and mutual belief are linked; but unlike the BDI account they are *not* equivalent. Where $D = \{x, y\}$, the **proof** that $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} MB_{x,y}\phi$ is as follows:

1. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_x\phi$ — Ax2, Sincerity
2. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_y\phi$ — Ax2, Sincerity
3. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_y\mathcal{B}_x\phi$ — 1; CL is mutually believed
4. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}_x\phi)$ — 2, 3; $\mathcal{B}$ is KD45
5. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_x\mathcal{B}_y\phi$ — 2; CL is mutually believed
6. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_x(\phi \wedge \mathcal{B}_y\phi)$ — 1, 5; $\mathcal{B}$ is KD45
7. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} MB_{x,y}\phi$ — 4,6; Induction Scheme

□

Thus grounded content is normally mutually believed; e.g., it is in (2) and (3), but not in (1).

Announcements affect intentions as well as beliefs. For instance, an intuitively compelling axiom is `Intent to Ground`: if a commits to ϕ, then normally he commits that he intends (written $\mathcal{I}_a$) that his interlocutors commit to it too, if they haven't done so already. A version of `Sincerity` also applies to intentions, and like `Sincerity` for beliefs requires adding an action operator $\sharp_a$ with a similar interpretation to $\flat_a$, to effect a model transition for the update of intentions.

- `Intent to Ground`:
  $(b \in D \wedge \mathcal{P}_{a,D}\phi \wedge \neg\mathcal{P}_{b,D}\phi) > \mathcal{P}_{a,D}\mathcal{I}_a\mathcal{P}_{b,D}\phi$
- `Sincerity on Intentions`:
  $\mathcal{P}_{a,D}\mathcal{I}_a\phi > \sharp_a\phi$

Together with axioms that link various speech act types to their illocutionary purpose and an axiom of `Cooperativity` ($\mathcal{P}_{a,D}\mathcal{I}_a\phi > \mathcal{I}_b\phi$; see below), these axioms ensure that the intentions behind a's current announcement become by default the intentions of all agents in D. Thus what one agent says can affect another agent's subsequent behaviour. For instance, the axioms predict from (1a) that A intends B to commit to C is stupid; B does this by announcing (1b). The axioms also predict from (1c) that A intends C to commit to C is not stupid, but A's intentions regarding B are more complex. A may not intend that B commit to (1c), and `Intent to Ground`, being defeasible, is compatible with this.

3.1 Desires

We have linked dialogue content to public commitment and the latter to belief and intention. But dialogue influences and is influenced by desires as well, and practical reasoning suggests that intentions are a byproduct of desires and beliefs. More precisely, rational agents intend those actions that maximise *expected utility*—utility reflecting one's desires or preferences, and *expectations* being based on *beliefs* about future outcomes. Preferences are thus distinct from but related to intentions.[1] We now address how an agent's preferences interact with other attitudes and dialogue content.

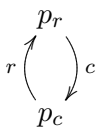
Games are a powerful model of preferences and actions among interacting agents. A game consists of a set of players and a set of strategies. Each strategy has a real-valued payoff or utility for each player. Typically the payoff for an individual is a function of *each* players' strategy, and intuitively, the payoff reflects that individual's preferences. A *Nash Equilibrium* (NE) is a combination of strategies that is optimal in that no player has a reason to deviate unilaterally from it. Games thus provide a method for computing one's next move in the dialogue. We illustrate this with a simple dialogue game in Table 4—a much simpler game than the ones that would underly the production of dialogues (1) to (3). In Table 4, R(ow) and C(olumn) are considering putdown moves (P_R and P_C) vs. praising moves. The cells indicate the utilities for agents R and C respectively for each combination of moves (e.g., column 2 row 2 defines the utilities for R and C when R praises C and C praises R). Note how the utilities for R and for C are influenced by what *both* agents do.

Since all utilities are defined, the game describes

[1] Preferences also have different logical properties: they can persist even after being realised while intentions don't; and they can be contrary to fact (one can prefer to be skiing right now while actually being at a meeting).

2/1	P_C	$\neg P_C$
P_R	$0, 0$	$3, -3$
$\neg P_R$	$-3, 3$	$4, 4$

Table 4: Simple Coordination Game



Preferences for C:

$\neg p_r \succ_c p_r$

p_r: $p_c \succ_c \neg p_c$

$\neg p_r$: $\neg p_c \succ_c p_c$

Preferences for R:

$\neg p_c \succ_r p_c$

p_c: $p_r \succ_r \neg p_r$

$\neg p_c$: $\neg p_r \succ_r p_r$

Figure 2: The CP-net for Table 4's 'Put Down' Game.

the complete preferences of each play with respect to all strategies. The two NEs are $(\neg P_R, \neg P_C)$ and (P_R, P_C). Utilities must be real values—standard game theory provides calculations of expected utility that combine probabilities over actions with the preferences for each player. But this sort of calculation is far too complex to be part of CL, which is a shallow logic for rough and ready decisions about discourse moves. To maintain a computationally effective CL, we need a *simpler* model of strategic reasoning that nevertheless *approximates* the types of interactions between expected moves and utility that game theory addresses.

Computationally efficient representations for strategic reasoning already exist. ***CP-nets*** (Boutilier et al., 2004) provide one such (qualitative) model for Boolean games (Bonzon, 2007)—games where like Table 4 each player controls propositional variables which he or she can make true or false (think of these as descriptions of actions that the agent performs, or not). A CP-net is designed to exploit the independence among the various conditions that affect an agent's preferences. It has two components: a directed *conditional preference graph* (CPG), which defines for each feature F its set of parent features $P(F)$ 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 parent values from $P(F)$.

For example, the CP-net for the 'put down' game from Table 4 is shown in Figure 2. p_c stands for C doing a put down move; similarly for p_r. The dependencies among features for each agent are shown with labelled arcs in the CPG. The CPT then distinguishes among the conditional preferences for agents R and C; e.g., $\neg p_r : \neg p_c \succ_c p_c$ stipulates that C prefers not to put down R rather than put him down, if R does not put down C. The semantics of CP-nets ensures that its conditional *ceteris paribus* preferences generate a total order $\succeq$ over all possible combinations of values of all features. Roughly put, the logic of CP-nets adheres to the following two (ranked) principles when generating this total order: 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 the total preference orderings for R and C for the CP-net in Figure 2 are as follows:

$(\neg p_r \wedge \neg p_c) \succ_c (\neg p_r \wedge p_c) \succ_c (p_r \wedge p_c) \succ_c (p_r \wedge \neg p_c)$

$(\neg p_r \wedge \neg p_c) \succ_r (p_r \wedge \neg p_c) \succ_r (p_r \wedge p_c) \succ_r (\neg p_r \wedge p_c)$

In line with the game in Table 4, these orderings yield two NEs: $(\neg p_r \wedge \neg p_c)$ and $(p_r \wedge p_c)$. While there are games whose CP-net representations do not validate *all* the game's NEs, Bonzon (2007) shows that CP-nets predict all NE when quite general conditions on the games are met.

Unfortunately, it is an inescapable fact that the preferences of other agents are hidden to us: one estimates them from their actions, including their utterances. CL must therefore use information from the dialogue to infer the CP-net for agents; CL must also make use of partial or underspecified CP-nets. For instance, what R knows about C and *vice versa* will determine how they should 'play' the 'Put down' game. If R has the preferences from Figure 2, but C is a jerk—in other words, his preference is to play a putdown move, *whatever* the circumstances (so in contrast to Figure 2, his CPG contains no dependencies on p_c and his CPT is simply $p_c \succ_c \neg p_c$)—then this revised CP-net has a different NE; namely, $p_r \wedge p_c$. So, using the general strategy that R should choose a future dialogue move according to NE, he will do p_r. If, on the other hand C is not a jerk, with the CP-net from Figure 2, then R should play $\neg p_r$. So if R doesn't know if C is a jerk or a non-jerk, he can't guarantee his next move to be optimal. Such put-down games might therefore be useful for establishing what sort of person one is dealing with. R might engage in this game to

see how C acts (is C a jerk, or a non-jerk?), before R makes conversational moves towards other ends where the penalties are much higher.

3.2 Back to Cognitive Logic

As shown in Lang et al. (2003), one can translate CP-nets into a conditional logic. We can do the same with the weak conditional $>$ from CL. Our representation of a conditional preference in terms of $>$ introduces a predicate OK that labels a world as being a good outcome (Asher and Bonevac, 2005), where OK is always strictly preferred to $\neg OK$. We then adopt the following definition of agent a's conditional preference $\phi : \psi \succ_a \neg\psi$:

- `Preference in CL:` $(\phi : \psi \succ_a \neg\psi) \Leftrightarrow$
  $\phi \rightarrow (\neg((\phi \wedge \psi) > \neg OK_a) \wedge$
  $((\phi \wedge \neg\psi) > \neg OK_a))$

In words, some normal $\phi \wedge \psi$ worlds are better than all normal $\phi \wedge \neg\psi$ worlds. The unconditional preference $\psi \succ_a \neg\psi$ is thus $\neg(\psi > \neg OK_a) \wedge (\neg\psi > \neg OK_a)$. In contrast to reasoning with games and CP-nets directly, `Preference in CL` allows CL to reason with *partial* information about the relative preferences among all possible actions.

Let's now investigate how preferences link to other attitudes. First, there is a rationality constraint linking preferences to intentions. Consider an unconditional preference first:

- `Preferences to Intentions:`
  $(\phi \succ_a \neg\phi \wedge \mathcal{B}_a \Diamond_G \phi) > \sharp_a \phi$

In words, if an agent, all things considered, prefers ϕ and believes there to be a strategy for achieving ϕ in the contextually supplied game or decision problem G (our gloss for $\Diamond_G$), then defeasibly he forms the intention to ϕ. Preferences within a game allow us with `Preferences to Intentions` to specify a version of what Asher and Lascarides (2003) call the *Practical Syllogism* (PS), which links beliefs, intentions and the *choice* that marks one's preferred way of achieving goals.[2] Suppose G has a unique optimal solution s for agent a such that $s > \phi$. Then a prefers the sequence of moves leading to s to any alternative sequence, and by `Preferences to Intentions` that sequence is intended. Asher and Lascarides (2003) used PS to infer an agent's beliefs and intentions from his behaviour and *vice versa*. We can now do this without PS as a separate principle.

On the other hand, when speakers *publicly commit* to a certain intention or to a preference, then this is an at least defeasible sign about their *actual* preferences. So when reasoning about an agent, if he commits to a certain intention or a certain preference, this licenses a dynamic update of one's model of his preferences ($\heartsuit$ is the 'preferences' action operator, where $\heartsuit_a \chi$ effects a model transition where conditional preference χ is added to a's preferences, so long as it is consistent to do so):

- `Commitments to Preferences:`
  $(\mathcal{P}_{a,D}\mathcal{I}_a\phi \vee [\mathcal{P}_{a,D}(\phi \succ_a \neg\phi)]) >$
  $\heartsuit_a(\phi \succ_a \neg\phi)$

In cooperative games, it seems reasonable to suppose that in general if one agent prefers a certain outcome then so does another. That is, $(\phi \succ_a \psi) > (\phi \succ_b \psi)$ for players a, b in a cooperative game. This allows us together with `Preferences to Intentions` and `Commitments to Preferences` to derive the follow Cooperativity axiom:

- `Cooperativity:` $\mathcal{P}_{a,D}\mathcal{I}_a\phi > \mathcal{I}_b\phi$

Thus by using CP-nets and their translation into CL, we can deepen the foundations of CL itself, rendering more transparent the axioms assumed there.

We can also now make dynamic the interaction between information about cognitive states and dialogue moves. For example, let's examine R and C playing the putdown game in three scenarios that vary on how partial (or complete) R's and C's knowledge of each other's preferences are. First, suppose R and C have complete (and accurate) knowledge of each others preferences, which are those in Figure 2. Then by `Preferences to Intentions` R will intend $\neg p_r$ (i.e., praise C), and similarly C will intend $\neg p_c$ (i.e., praise R). By `Intent to Ground` both intentions will become also mutual intentions of R and C. And both have a rational expectation for how the verbal exchange

[2] They state PS as follows:
$(\mathcal{I}_a(\psi) \wedge \mathcal{B}_a((\phi > \psi) \wedge \mathit{choice}_a(\phi, \psi)) > \mathcal{I}_a(\phi)$
In words, if a intends that ψ, and he believes that ϕ normally leads to ψ and moreover ϕ is a's choice for achieving ψ, then normally a intends that ϕ. By treating the relation choice_a as primitive, the CL lacked the reasoning that agents engage in for finding optimal ways of achieving goals. We remedy this here.

will go.

Now consider the case where R's preferences are those in Figure 2 but R does not know if C is a jerk or not. On the other hand, C believes his own and R's preferences to be those given in Figure 2. Then R may not yet have formed an intention with respect to the goal, since he has no information on C's preferences or intentions. But C will act as above and thus R will learn about C's actual intentions. That is, on observing C perform $\neg p_c$ R will know that C intended it,[3] and by `Commitments to Preferences` she will update her model of C's preferences with $\neg p_c \succ_c p_c$. This now allows her to use the CP-net so-constructed to make the move that maximises her preferences—i.e., $\neg p_r$.

Finally, consider the case where R and C meet for the first time and don't know anything about each other's preferences. If R is to make the first move, then unlike the prior case R cannot use C's actions to influence her move. Instead, she must reason by 'cases', using each CP-net that is compatible with her own preferences. Suppose that R's preferences are those in Figure 2, and furthermore, R knows C to be either a non-jerk (as in Figure 2) or a jerk (making C's CP-net simply $p_c \succ_c \neg p_c$). Then R can reason as follows. If C is a non-jerk, then C prefers $\neg p_c$ on condition that R performs a $\neg p_r$ (reasoning as before), making R's best move $\neg p_r$. On the other hand, if C is a jerk, then C prefers p_c regardless, making R's best move p_r. R would therefore require further strategies for deciding which of p_r vs. $\neg p_r$ to prefer. For instance, R might 'hope for the best' and perform $\neg p_r$. In any case, where all that is involved is an insult, R may consider it better to potentially receive an insult and know about C's desires than to behave like a jerk herself. An extension of the CP-net could model these additional preferences.

4 Conclusions

In this paper we developed a cognitive logic for discourse interpretation that extends dynamic logics of public announcement. The extensions provide default links between public announcements and cognitive attitudes. It validates that grounding normally leads to mutual belief, but not always (see (1)). We also argued for representing preferences as $>$-statements, and highlighted the relationship between this and CP-nets—a compact way of representing Boolean games of the kind that have been used to model dialogue strategies. We thus linked within CL game-theoretic principles to general axioms of rationality and cooperativity. This affords a 'generate-and-test' way of deciding one's next dialogue move, even when one has only partial information about another agent's preferences. In future work, we plan to explore how to use this CL to model calculable implicatures (Grice, 1975).

[3] See Asher and Lascarides (2003) for details.

References

L. Amgoud. A formal framework for handling conflicting desires. In *Proceedings of ECSQARU*, 2003.

N. Asher and D. Bonevac. Free choice permission is strong permission. *Synthese*, pages 22–43, 2005.

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

A. Baltag, L.S. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicions. Technical Report SEN-R9922, Centrum voor Wiskunde en Informatica, 1999.

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. *JAIR*, 21:135–191, 2004.

B. Gaudou, A. Herzig, and D. Longin. Grounding and the expression of belief. 2006.

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

B. Grosz and C. Sidner. Plans for discourse. In J. Morgan P. R. Cohen and M. Pollack, editors, *Intentions in Communication*, pages 365–388. MIT Press, 1990.

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

J. Lang, L. van der Torre, and E. Weydert. Hidden uncertainty in the logical representation of desires. In *Proceedings IJCAI*, pages 685–690, 2003.

A. Lascarides and N. Asher. Agreements and disputes in dialogue. Proceedings of SIGDIAL, 2008.

D. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.

Dialogue-Grammar Correspondence in Dynamic Syntax

Andrew Gargett[†], **Eleni Gregoromichelaki**[†], **Chris Howes**[‡], **Yo Sato**[‡]
[†]King's College London, Strand, London WC2R 2LS, UK
{andrew.gargett,eleni.gregor}@kcl.ac.uk
[‡]Queen Mary University of London, Mile End Road, London E1 4NS, UK
{chrizba,yosato}@dcs.qmul.ac.uk

Abstract

In this paper, we argue, contra a prevailing trend to classify elliptical structures in terms of sub-types specific to conversational dialogue, that despite their diversity of uses in conversational dialogue, such fragments are analysable in terms of structure-building mechanisms that have motivation elsewhere in the grammar (the framework adopted is *Dynamic Syntax*, Kempson et al. (2001); Cann et al. (2005)). Fragment types modelled include *reformulations*, *clarification requests*, *extensions*, *corrections* and *acknowledgements*. We argue that incremental use of such ellipses serves a specific role in dialogue, namely a means of incrementally narrowing down the range of otherwise mushrooming alternative structural and interpretative options, a problem known to constitute a major challenge to any parsing system. We conclude that with grammar seen as a set of parse procedures, we have a basis for an integrated characterisation of dialogue phenomena while nonetheless not defining a grammar of conversational dialogue.

1 Introduction

In confronting the challenge of providing formal models of dialogue, with its plethora of fragments and rich variation in modes of context-dependent construal, it might seem that linguists face two types of methodological choice. Either (a) conversational dialogue demonstrates dialogue-specific mechanisms, for which a grammar specific to such activity must be constructed; or (b) variation arises due to the employment of independent parsing/production systems which are nevertheless based on some mode-neutral grammar. However, as dialogue research continues to develop, there are intermediate possibilities, and in this paper we discuss the approach developed within *Dynamic Syntax* (*DS*, Kempson et al. 2001, Cann et al. 2005), a grammar framework within which, not only the parser, but indeed "syntax" itself is seen as the progressive construction of semantic representations set in context. Here we extend the analyses presented in Kempson et al. (2007) to a range of further fragment types, in particular *reformulations*, *fragment requests* and *corrections* accompanied by *extensions*. From a DS perspective, such apparently dialogue-specific constructions can be seen to result from perfectly general structural processes, despite being characteristic of cross-party conversational data.

Further, we claim that the grammar itself constitutes the basis for parsing strategies that facilitate an efficient online processing, both structural and semantic. In this respect, the DS dialogue model provides the means of achieving this *during* the course of the sub-sentential construction process, demonstrating that timely application of such generally available "syntactic" mechanisms directly contributes to the human processor's high degree of success in linguistic interaction. Contrary to conventional assumptions of the grammar-parser feeding relation whereby the parser exclusively handles disambiguation, we conclude that grammars, as employed in dialogue, can also be seen as restricting ambiguity provided their formal specification can model this incremental facilitating function.

2 Background

The data we focus on are non-repetitive fragment forms of acknowledgements, clarifications and corrections (henceforth, A female, B male):

(1) A: Bob left.
B: (Yeah,) the accounts guy.

(2) A: They X-rayed me, and took
a urine sample, took a blood sample.
A: Er, the doctor
B: Chorlton?

A: Chorlton, mhm, he examined me, erm, he, he said now they were on about a slight [shadow] on my heart. [BNC: KPY 1005-1008]

(3) A: Are you left or
B: Right-handed.

(4) A: Bob left.
B: Rob?
A: (No,) (Bob,) the accounts guy.

Even though in the literature the NP fragments in (2) - (4) might be characterised as distinct constructions, they all illustrate how speakers and hearers may contribute, in some sense to be made precise, to the joint enterprise of establishing some shared communicative content, in what might be loosely called *split utterances*. And even (1), an *acknowledgement*, can be seen this way upon analysis: B's addition is similar to an afterthought *extension* added to A's fully sentential utterance. It can be seen in (2) that such joint construction of content can proceed incrementally: the *clarification request* in the form of a *reformulation* is provided by B and resolved by A within the construction of a single proposition. The attested example in (3) represents an intermediate case, in which the respondent realising what the question is provides the answer AS the *completion* of the initiator's question, so that the fragment serves simultaneously as question and answer. In (4) the fragment reply involves *correction*, with parties to the conversation confronting the need for negotiation as to whose information is more reliable before coordination can be said to be achieved. Nevertheless such corrections can be also *extensions* in the above sense, enabling a single conjoined propositional content to be derived before the requisite coordination can be achieved.

It might seem that such illustration of diversity of fragment uses is ample evidence of the need for conversation-specific rules to be articulated as part of a grammar. Indeed, Fernández (2006) presents a thorough taxonomy, as well as detailed formal and computational modelling of *Non-sentential Utterances* (NSUs), referring to contributions such as (1) as *repeated acknowledgements* involving *reformulation*. Since such fragments require contextual information singling out a particular constituent of the previous utterance, Fernández models such constructions via type-specific "accommodation rules" which make a constituent of the antecedent utterance "topical". The semantic effect of acknowledgement is then derived by applying an appropriately defined utterance type for such fragments to the newly constructed context. A distinct form of contextual accommodation is employed to model so-called *helpful rejection* fragments, as in (4) (without the reformulation), whereby a *wh*-question is accommodated in the context by abstracting over the content of one of the sub-constituents of the previous utterance. The content of the rejection is derived by applying this *wh*-question in the context to the content of the fragment (see also Schlangen (2003) for another classification and analysis).

The alternative explored here is whether phenomena such as (1)-(2), both of which are non-repetitive appositional next-speaker contributions, can be handled uniformly using the mechanisms for structure-building made available in the core grammar, without recourse to conversation-specific extensions of that grammar and contextual accommodation rules. The range of interpretations these fragments receive in actual dialogue seem to form continua with no well-defined boundaries and mixing of functions (see (3)-(4) and comments in Schlangen (2003)). Thus we propose that the grammar itself simply provides mechanisms for processing/integrating such fragments in the current structure while their precise contribution to the communicative interaction is either calculated by pragmatic inferencing (as in e.g. Schlangen (2003)) or, as seems most often to be the case, left underspecified. The framework within which the explanation will be provided is Dynamic Syntax, in which the dynamics of how information accrues in language processing is the core of the syntactic explanation.

One bonus of the stance taken here is the promise it offers for elucidating the grammar-parser contribution to the disambiguation task. Part of the challenge of modelling dialogue is the apparent multiplicity of interpretive and structural options opened up during processing by the recurrent, often overlapping fragments as seen in (2) above. Thus, it might seem that the rich array of elliptical fragments available in dialogue adds to the complexity of the interpretive task, owing to their high degree of context-dependence (hence the need for accommodation and construction-specific interpretation rules). However, an alternative point of view is to see such phenomena as providing a window on how interlocutors exploit the

incrementality afforded by the grammar to manage the explosion of interpretative/structural options multiplying at each step. The context-dependent interpretation of fragments, when employed incrementally, enables the hearer to immediately respond to a previous utterance at any relevant point in the construction process, thereby enabling interlocutors to (incrementally) constrain interpretation during the very process in which it is developed.

Modelling this kind of flexibility in processing requires fine-grained control of how the current utterance can be combined with previous contextual information. Grammatical frameworks which take the radical context dependency of linguistic processing as being outside the remit of the grammar might make it seem that these phenomena require distinct mechanisms. Alternatively, however, the tight coordination of parsing and generation as defined in the *Dynamic Syntax* model of dialogue (Purver et al. (2006)) enables a straightforward account of how the context-dependence of both tasks allows participants to economise on processing.

3 Dynamic Syntax: A Sketch

Dynamic Syntax (*DS*) is a parsing-based approach to linguistic modelling, involving strictly sequential interpretation of linguistic strings. The model is implemented via goal-directed growth of tree structures and their decorations formalised using *LOFT* (Blackburn and Meyer-Viol (1994)), with modal operators $\langle\uparrow\rangle, \langle\downarrow\rangle$ to define concepts of *mother* and *daughter*, and their iterated counterparts, $\langle\uparrow_*\rangle, \langle\downarrow_*\rangle$, to define the notions *be dominated by* and *dominate*. *Underspecification* and update are core aspects of the grammar itself and involve strictly monotonic information growth for any dimension of tree structures and decorations. Underspecification is employed at all levels of tree relations (mother, daughter etc.), as well as formulae and type values, each having an associated *requirement* that drives the goal-directed process of update. For example, an underspecified subject node of a tree may have a requirement expressed in DS with the node decoration $?Ty(e)$, for which the only legitimate updates are logical expressions of individual type ($Ty(e)$); but requirements may also take a modal form, e.g. $?\langle\uparrow\rangle Ty(e \rightarrow t)$, a restriction that the mother node be decorated with a formula of predicate type. Requirements are essential to the dynamics informing the DS account: all requirements must be satisfied if the construction process is to lead to a successful outcome.

Structure is built from lexical and general computational actions. *Computational actions* govern general tree constructional processes, such as introducing and updating structure, as well as compiling interpretation for all non-terminal nodes in the tree, once individual leaf nodes are successfully decorated (with no outstanding requirements). This may include construction of only weakly specified tree relations, characterised only as dominated by some node from which they are constructed (*unfixed nodes*), with subsequent update (unlike van Leusen and Muskens (2003), partial trees are part of the model). Individual lexical items also provide procedures for building structure in the form of *lexical actions*, expressed in exactly the same terms as the more general processes, inducing both nodes and decorations. Thus *partial trees* grow incrementally driven by procedures associated with particular words as they are encountered, with a *pointer*, $\Diamond$, recording the parser's progress.

Complete individual trees are taken to correspond to predicate-argument structures. More complex structures can be obtained via a general tree adjunction operation defined to license the construction of a tree sharing some term with another newly constructed tree, yielding so-called *Linked Trees* (Kempson et al. 2001). The resulting combined information from the adjoined trees is modelled as a conjunction of terms at the node *from* which the link is made. Importantly, adjunction, as other forms of construction and update, can be employed to model how subsequent speakers may dynamically provide fragmentary extensions in response to the previous utterance.

Structural as well as content underspecification play important roles in facilitating successful linguistic interaction. The content underspecification of pronouns is represented as a place-holding metavariable, noted as e.g. $\mathbf{U}$, plus an associated requirement for update by an appropriate term value: $?\exists\mathbf{x}.Fo(\mathbf{x})$. Similarly, *names* are represented as initially introducing place-holders associated with a constraint providing the name of the individual entity picked out. For example, the name *Bill* contributes the decoration $U_{Bill'(\mathbf{U})}, Ty(e)$. The subscript specification is shorthand for a transition across a LINK relation to a tree whose top node is decorated with a formula $Bill'(\mathbf{U})$, the name being taken as a predicate or name specification of individuals thus

restricting possible updates to the metavariable[1]. Names can thus be seen as a procedure for identifying the individual being talked about, with a logical constant (e.g. $m21$, $m23$ etc. picking out uniquely this individual) eventually replacing the metavariable on the emergent tree. According to the DS account, the update of metavariables can be accomplished if the context contains an appropriate term for substitution. *Context* in DS involves storage of *parse states*, i.e., the storing of partial tree, word sequence to date, plus the actions used in building up the partial tree.

A major aspect of the DS dialogue model is that both *generation* and *parsing* are goal-directed and INCREMENTAL, with parsing as the underlying mechanism and generation parasitic on it. A hearer builds a succession of partial parse trees in order to achieve an interpretation of the speaker's message. A speaker is modelled in DS as doing exactly the same only (s)he also has available a *goal tree* representing what they wish to say. Each possible step in generation, an utterance of a word, is governed by whatever step is licensed by the parsing formalism, constrained via the required *subsumption* relation of the goal tree by the thus-far constructed "parse" (partial) tree. By updating their growing "parse" tree relative to the goal tree (via a combination of incremental parsing and lexical search), speakers produce the associated natural language string.

The DS model of dialogue requires defining and taking into account both the speaker's goal and parse trees, as well as the hearer's parse tree. For fragment construal, we are interested in the extent to which B has successfully parsed what A has said, with the ability at any stage to interrupt to ask for clarification, reformulate, or provide a correction, by either repeating the expression or producing an alternative. As we shall see, B's parse tree reveals where need of clarification or miscommunication occurs, as it will be at that node from which a sub-routine extending it takes place. According to the DS model of generation, repeating or extending a constituent of A's utterance is licensed only if B's goal tree matches or extends a parse tree updated with the relevant subpart of A's utterance. Indeed, this update is what B is seeking to clarify, correct or acknowledge.

Notice that because of the incremental definition of DS, B can reuse the already constructed (partial) parse tree in their context, thereby starting at this point, rather than having to rebuild an entire propositional tree or subtree (e.g. of type e). Exploiting the assumed parity of representations in this way enables hearers to provide immediate feedback to the previous speaker, the effect being to narrow the focus on particular aspects of the interpretive space. The advantage of this emerges in the unified characterisation of any type of *ellipsis* construal as strictly context-dependence. Since context in DS involves the storing of current partial tree, word sequence to date, plus the actions used to date to build the partial tree, ellipsis construal can target any of those stored elements. In particular, for split/joint utterances, this enables switch from hearer to speaker at any arbitrary point in the dialogue, without such fragments having to be interpreted as propositional in type (as is standard elsewhere, e.g. Purver (2004)).[2] This can then capture the dynamics involved in taking what the other speaker has just uttered, with the potential at any point to update it to accord with one's own emerging understanding of the interaction. In this way, speakers are able to guide each other's interpretations, and thus *jointly* narrow down as early as possible the burgeoning interpretive space.

4 NSU fragments in Dynamic Syntax

4.1 Non-repetitive Acknowledgement

From a DS perspective, phenomena like *reformulations* as in (1), or *extensions* to what one understands of the other speaker's utterance, (2), can be handled with exactly the same mechanisms as the sentence-internal phenomenon independently identifiable as *apposition* and illustrated below:

(5) A friend of my mother's, someone very famous, is coming to stay.

(6) Bob, the friend of Ruth's, is coming to stay.

According to Cann et al. (2005), such structures are analysed as involving the building of paired terms across a tree transition, building *linked* structures defined to share a term. Reflecting this constraint, the update rule for such structures then takes the pair of type e terms so formed and yields

[1] These *linked* structures are suppressed in all diagrams.

[2] Given the DS concept of linked trees projecting propositional content, we anticipate that this mechanism will be extendable to fragment construal involving inference (see e.g. Schlangen (2003), Schlangen and Lascarides (2003))

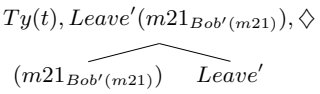
a term whose compound restrictor is made up of the predicative content from each.

We now have the basis for analysing extensions and non-repetitive acknowledgements which build on what has been previously said by way of confirming the previous utterance. Recall examples (1) and (2). There are two ways for the processing of fragments which reformulate an interlocutor A's utterance: either (a) as interruptions of her, A's, utterance in which case immediate confirmation of identification of the individual concerned is provided, see (2), or (b) as confirmations/extensions of A's utterance after the whole of her utterance has been integrated, see (1). Both are modelled by DS as incremental additions.

Turning to (1), B's response *(Yeah,) the accounts guy* constitutes both a reformulation of A's utterance, as well as an extension of A's referring expression, having the same effect as processing the appositive expression *Bob, the accounts guy*. This means that B has processed A's original utterance, according to some identification of the individual associated with the name *Bob*: that is to say, they have constructed a full content representation for this utterance. B's reformulation has the effect of acknowledgement because it signals to A that he has processed/understood her asserted content, and, moreover, has no objection to the content, *unless* mistaken in that identification.

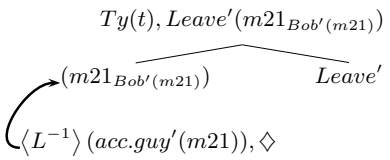
In DS terms, B's context consists of the following tree after processing A's utterance:

(7) B's Context for *Yeah*[3]



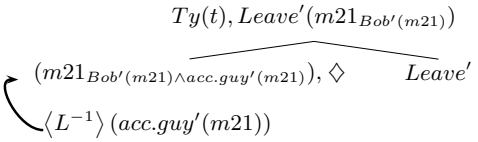
It is now open to B to re-use this representation, stored in his context, as the point of departure for generating the expression *the accounts guy*. In this case his own goal tree will now be decorated with a composite term made up both from the term recovered from parsing A's utterance and the new addition. Simplistically, all this requires is attaching a *linked* tree to the correct node, and then processing the content of the apposition in order to produce the words required. The defined steps include shifting the pointer to the appropriate node, projection of a *linked* tree from that node and processing the words *the accounts guy* (the *linked* tree is simplified below):

(8) B's "parse" tree licensing production of *the accounts guy*: LINK adjunction



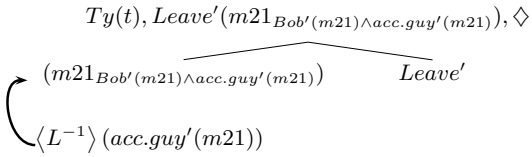
Updating this representation according to the DS processing protocol involves adding the acquired restrictions at the node from which the *linked* tree is projected (individual stages here suppressed):

(9) Updating B's "parse" tree licensing production of *the accounts guy*



Finally, the information is passed up to the top node of the main tree, completing the parse tree to match B's goal tree in uttering the expression *the accounts guy* :

(10) Completing B's "parse" tree licensing production of *the accounts guy*



4.2 Non-repetitive Clarification

In the acknowledgement case above, the term relative to which the *linked* structure is built is fixed; but the very same mechanism can be used when the interlocutor needs clarification. In (2), B again takes as his goal tree a tree decorated with an expansion of the term constructed from parsing A's utterance but nevertheless picking out the same individual. Using the very same mechanism as in (1) of building a *linked* structure constrained to induce shared terms, B provides a distinct expression, the name *Chorlton*, this time before he has completed the parse tree for A's utterance. This name, contributing a metavariable plus the constraint that the individual picked out must be named *Chorlton*, is used to decorate the

[3]Words like *yeah* and *no* are analysed as discourse markers which do not contribute truth conditional content, hence are not represented on the trees

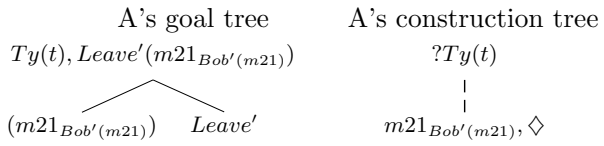


Figure 1: Licensing production of a correction by *ADJUNCTION

linked node so that it makes explicit the additional predicative constraint on the individual being described. The outcome of this process, when the linked structure is evaluated, is a composite term $m21_{Doctor'(m21) \wedge Chorlton'(m21)}$. This process, therefore, is identical to that employed in B's utterance in (1), though to rather different effect at this intermediate stage in the interpretation process. This extension of the term is confirmed by A, this time trivially replicating the composite term which processing B's utterance has led to (see Kempson et al 2007 for such trivial goal tree-parse tree matches). The eventual effect of the process of inducing *linked* structures to be decorated by coreferential type e terms may thus vary across monologue and different dialogue applications, yielding different interpretations, but the mechanism is the same.

4.3 Correction

It might be argued nonetheless that correction is intrinsically a dialogue phenomenon. In (4) for example, reproduced below:

(4) A: Bob left.
B: Rob?
A: (No,) (Bob,) the accounts guy.

As one alternative, we assume here that B has misheard and requests confirmation of what he has perceived A as saying. A in turn rejects B's utterance and provides more information. Presuming rejection as simple disagreement (i.e. the utterance has been understood, but judged as incorrect), in DS terms, this means that A has in mind a goal tree that licensed what she had produced, which is distinct from the one derived by processing B's clarification. As shown in Kempson et al. (2007), this means that A has been unable to process B's clarification request as an extension of her own context. Instead she can parse the clarification by exploiting the potential for introducing an initially structurally underspecified tree-node to accommodate the contribution of the word *Rob*. Subsequently, by re-running the actions stored in context previously by processing her own utterance of the word *left*, she is able to complete the integration of the fragment.

In order to produce the following correction, A is required to establish as the current most recent representation in context her original goal tree. This can be monotonically achieved by recovering and copying this original goal tree to serve as the current most immediate context[4]. Under these circumstances, given the DS grammar-as-parser perspective, several strategies are now available. A is licensed to repeat the name *Bob* by locally extending the node in the context tree where the representation of the individual referred to is located by using the rule of LATE*ADJUNCTION, a process which involves building a node of type e from a dominating node of that type (illustrated in Kempson et al. 2007). An alternative way of licensing repetition of the word *Bob* is to employ one of the strategies generally available for the parsing of long distance dependencies i.e. constructing initial tree nodes as unfixed (*ADJUNCTION).

Starting with Fig 1 above, illustrating the introduction of the unfixed node, we show here how the latter strategy can be exploited to license the production of the fragment. An option available to A at this point is to introduce, in addition or exclusively, a reformulation of her original utterance in order to facilitate identification of the named individual which proved problematic for B previously. She can answer B's utterance of *Rob* with *(No,) (Bob,) the accounts guy*, as in (4) or simply with *(No,) the accounts guy*. Both are licensed by the DS parsing mechanism without more ado. The structure[5] derived by processing such an extension is exactly that of (1) above (compare goal tree in Fig 2 and tree in (10)). As mentioned before, *context*, as defined in DS, keeps track not only of tree representations and words but also of actions contributed by the words and utilised in building up the tree representations. Production of

[4]Corrected representations must be maintained in the context as they can provide antecedents for subsequent anaphora.

[5]Note that DS trees represent derived content rather than structure over natural language strings.

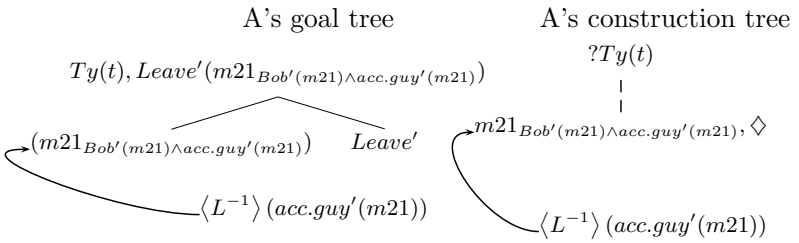


Figure 2: LINK ADJUNCTION and checking goal tree subsumption

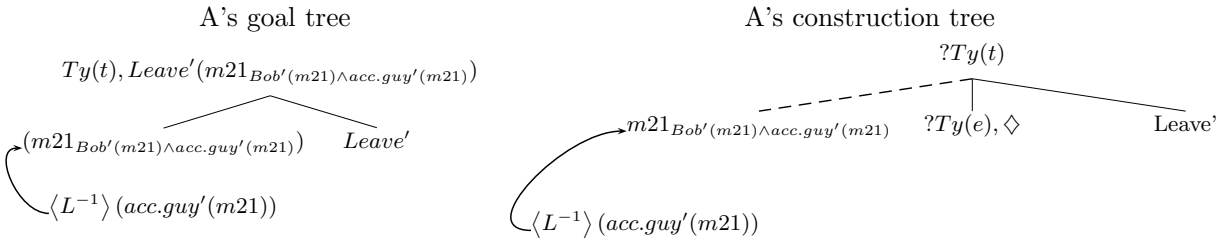


Figure 3: Retrieving and rerunning the actions for *left*, pointer return to subject node and checking goal tree subsumption

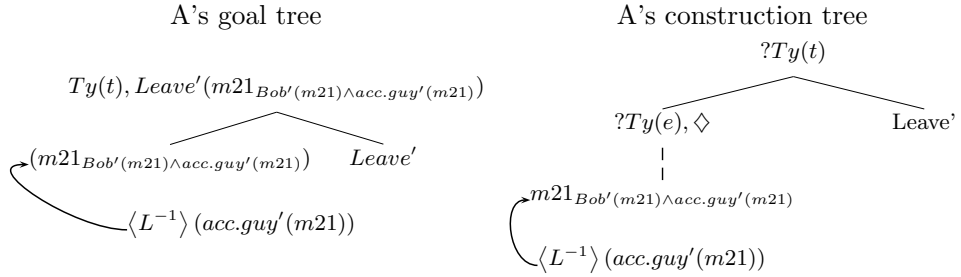


Figure 4: Preparation for UNIFICATION and checking goal tree subsumption

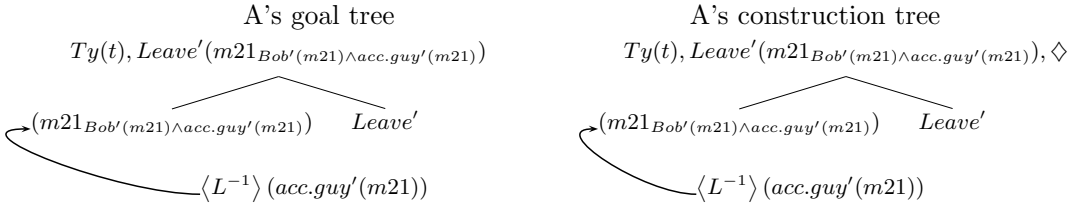


Figure 5: Licensing the production of correction and extension: completed tree matching the goal tree

the correction in (4) is licensed to be fragmental because the original actions for parsing/producing the word *left* are available in the context and can be recalled to complete the structure initiated by processing/producing the name *Bob* (see Fig 3-5).

4.4 Structure and Dialogue Function

In the examples considered so far, we have seen how a single type of mechanism can serve distinct functions. A more striking case is (3), where the hearer, B, is able to leap to a hypothesis as to how A's question is going to be completed, and provides that completion by way of answer. Here we have the case where more than one function can be fulfilled even by a single utterance. As in (1)-(2), license for such a use turns on taking the context that was constructed by parsing input from the interlocutor as the point of departure. That B is extending the structure set up by A's utterance is self-evident; but in addition, both A's utterance, if she had completed it, and B's utterance, as presented, are elliptical as to the second disjunct. The success

of this particular form of split utterance turns on the fact that what A is presenting is a duplex *yes-no* question with both possible answers provided by the two disjuncts. So in completing it by providing just the second disjunct, B can succeed in answering the question while simultaneously completing it. Though there is more to say here, the significance of (3) lies in the use of the single expression *right-handed* to fulfil two functions, both the completion of a question and the provision of an answer. In DS this can be modelled, reflecting the phenomenon itself, without having to assume the superimposition of two distinct structures, one upon the other. Incidentally, this is a case contradicting what is supposedly unique to such interrupting completions, namely, that they require acknowledgement by the hearer before proceeding.

5 Conclusion

As these fragments and their construal show, despite serving distinct functions in dialogue, the mechanisms which make such diversity possible are general strategies for tree growth. In all cases, the advantage which use of fragments provides is a "least effort" means of re-employing previous content/structure/actions which constitute the *context*. As modelled in DS, it is more economical to reuse information from context rather than constructing representations afresh (via costly processes of lexical retrieval, choice of alternative parsing strategies, etc.).

A further quandary in dialogue construal is that, despite such avenues for economising their efforts, interlocutors are nevertheless faced with an increasing set of interpretative options at any point during the construction of representations. One option available to hearers is to delay a disambiguating move until further input potentially resolves the uncertainty. However, as further input is processed and parsing/interpretive options increase potentially rapidly, maintenance of these open options becomes difficult for a human processor. The incremental definition of the DS formalism allows for the modelling of an alternative available to hearers: at any point they could opt to intervene immediately and make a direct appeal to the speaker for more information at the maximally relevant point during construction. It seems clear that the latter would be a preferable strategy and this is what clause-medial fragment interruptions, (2), illustrate.

The phenomena examined here are also cases where speaker's and hearer's representations, despite attempts at coordination, may nevertheless separate sufficiently for them to have to seek to explicitly "repair" the communication (see especially (4)). In the model presented here, the dynamics of interaction allow fully incremental generation and integration of fragmental utterances so that interlocutors can be taken to constantly provide optimal evidence of each other's representations with necessary adjuncts being able to be incrementally introduced. Thus, fragment construal is here modelled sub-sententially with no lifting devices to yield a propositional unit as part of some putative discourse grammar. Indeed, no structures/strategies are posited specific to individual discourse functions to which a fragment is put.

Acknowledgements

We gratefully acknowledge extensive helpful discussions with Ruth Kempson and valuable comments from three anonymous reviewers. This work was supported by grants ESRC RES-062-23-0962 and Leverhulme F07 04OU.

References

Patrick Blackburn and Wilfried Meyer-Viol. Linguistics, logic and finite trees. *Bulletin of the IGPL*, 2:3–31, 1994.

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

Raquel Fernández. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. PhD thesis, King's College London, University of London, 2006.

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

Ruth Kempson, Andrew Gargett, and Eleni Gregoromichelaki. Clarification requests: An incremental account. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue (DECALOG)*, 2007.

Matthew Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, University of London, forthcoming 2004.

Matthew Purver, Ronnie Cann, and Ruth Kempson. Grammars as parsers: Meeting the dialogue challenge. *Research on Language and Computation*, 4(2-3):289–326, 2006.

David Schlangen. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. PhD thesis, University of Edinburgh, 2003.

David Schlangen and Alex Lascarides. The interpretation of non-sentential utterances in dialogue. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 62–71, Sapporo, Japan, July 2003. Association for Computational Linguistics.

Noor van Leusen and Reinhard Muskens. Construction by description in discourse representation. In J. Peregrin, editor, *Meaning: The Dynamic Turn*, chapter 12, pages 33–65. 2003.

User simulations for online adaptation and knowledge-alignment in Troubleshooting dialogue systems

Srinivasan Janarthanam
School of Informatics,
University of Edinburgh,
Edinburgh, EH8 9LW, GB.
s.janarthanam@ed.ac.uk

Oliver Lemon
School of Informatics,
University of Edinburgh,
Edinburgh, EH8 9LW, GB.
olemon@inf.ed.ac.uk

Abstract

We study the problem of alignment between dialogue participants, using the practical example of "troubleshooting" dialogue systems. Recent work on troubleshooting concerns automated spoken dialogue systems which support users who need to repair their internet connection. We address the problem that different users have different types of knowledge of problem domains, so that automated dialogue systems need to adapt online to the different knowledge of these users as it encounters them. We approach this problem using policy learning in a Markov Decision Process (MDP). In contrast to related work we propose a new user model which incorporates the different conceptual knowledge of different users, together with an environment simulation. We show that this model allows us to learn dialogue policies that automatically adapt online to new users, and that these policies are significantly better than threshold-based adaptive hand-coded policies for this problem.

1 Introduction

Adapting between conversation partners was first studied by Issacs & Clark (1987), where the partners identify each other's domain knowledge levels during conversation and share their knowledge, leading to task success. More recently, Larsson (2007) gives a formal account of how the meanings of NL expressions are adapted during conversations. These important aspects of dialogue are not addressed in current automated systems. Here, and in Lemon (2008), we propose a model that allows such decisions to be automatically optimized.

There has been much recent interest in automated dialogue systems for "troubleshooting" or "self-help". These cooperative systems are particularly interesting because they contain aspects of knowledge alignment and tutorial dialogue – for instance, some users may not know certain technical terms, so that systems must be able to "align" with their users, at least at a knowledge/concept-level. For example, some users may be happy with the term "ADSL filter", while others may need this explained to them before their problem can be solved. On the other hand, some users may be frustrated by unnecessary explanations. There is therefore an important tradeoff to be explored regarding how much additional explanation to provide to a particular user. Note that in general we will not know the knowledge profile of a user when they call the system, so our dialogue policies must be able to estimate the user type online, and continuously adapt their behavior based on the estimation.

All of this places additional requirements on our user models and user simulations for training these more complex systems. We provide a new user model for such purposes and show that it allows us to learn these types of adaptive dialogue policies.

2 Related work

Several user simulation models to support MDP -learning of dialogue management policies have been developed over recent years (Eckert et al 1997, Scheffler & Young 2001, Pietquin et al 2004, Georgila et al 2005, Cuayahuitl et al 2005, Schatzmann et al 2007). These simulations simulate users in travel planning and town-information domains.

They produce user responses based on various factors such as the system's action, user's goal/agenda, user's record of the dialogue, etc. Crucially, they only simulate a homogenous group of users, who always understand the system's actions completely and never ask for clarifications. The models also do not simulate the user's environment. In contrast, in this work we have focused on the user's domain knowledge and an environment simulation concerning troubleshooting systems.

Troubleshooting dialogue systems have been designed by Boye et al (2007), who presents a hand-coded system and Williams (2007) who uses machine learning. In Boye (2007) the task of fixing a broadband connection is hierarchically decomposed in to simpler tasks. When the user fails to respond, the system chooses alternative ways to solve the tasks. However, it always tries the standard procedure before choosing the alternatives. The dynamic tree structure that drives the conversation is interleaved with an "adapting-to-user" feature and is likely to become more complex to manually author in the case of realistic systems.

Williams (2007) presents a POMDP-based dialogue system for troubleshooting broadband connections. The system is trained to handle the uncertainty in user's observations and responses and provide the next appropriate instruction. Here the system learns *what* to ask or present rather than *how* to ask for information from the user. The system assumes a homogenous user population in terms of domain knowledge. However, in this work, we present an MDP system that learns to adapt to a user population where different users have different conceptual knowledge of the troubleshooting environment.

3 The Troubleshooting Domain & Dialogue Management

In this setup, the dialogue manager always directs the conversation, because (besides fixing their problem) the user does not have any other goal or agenda in order to direct the conversation. During the course of conversation, the dialogue manager asks the user to describe their troubleshooting environment (e.g. Modem lights etc). The information to be asked is handled using a hand-coded decision tree. The tree encodes what information to ask next based on the user's report on the environment so far. Previous work has shown that such trees can be learned from data (Williams 2007). While the decision tree decides what to ask next, the dialogue manager still has to decide *how* to ask for information and present instructions to users with different domain knowledge. Table 3.1 lists the dialogue manager actions related to the task at hand.

Table 3.1. Dialogue manager action set

1. Greet
2. Request_info
3. Extended_request_info
4. Request_action
5. Extended_request_action
6. Close_dialogue

In general, the dialogue manager must decide between a simple "request" act and an "extended request" act in order to request information from the user or to ask that they perform a manipulate action. An "extended request", although presented as a single turn in this dialogue act representation, is actually a sequence of system utterances that the system uses to educate the user about the concept that the system is querying/instructing about. For instance, a novice user may not know where the ADSL light is. In this case the system spends some time to inform the user where to look for this light. These extra utterances make an extended request more costly than a simple request (and this is later reflected in the reward function for learning this task).

The dialogue manager is also equipped with a *user estimation feature* that allows it to dynamically estimate the expertise of a user based on their responses and frustration. At the start of the session, it is assigned a default intermediate value of 5. This is later incremented or decremented based on evidence (e.g. whether the user answers the asked question). For an expert user, it increases from 5 and can reach 9 by the end of the dialogue. For a novice user, it can decrease to 0. However there is also uncertainty in user responses, since sometimes a novice user may be able to answer a question without extra help and hence may be misjudged as an expert. Similarly an expert user may fail to answer a question and be judged temporarily as a novice. But in the course of the conversation, the estimation function will usually correct itself as evidence about the user accumulates.

The dialogue manager's information state is composed of the following fields (table 3.2). There

are 8 binary slots and one 10-value integer slot in the dialogue information state, giving rise to a state space of 2560 (=2^8 * 10) states.

Table 3.2. Dialogue state features	
More_slots_to_ask	(binary)
Solution_found	(binary)
User_expertise_index	(0-9)
User_said_dont_know	(binary)
Modem_power_light_filled	(binary)
Adsl_light_filled	(binary)
Modem_filter_filled	(binary)
Phone_filter_filled	(binary)
Phone_line_working_filled	(binary)

4 Environment Simulation

This model simulates the troubleshooting environment around the user. The environment simulation represents all the objects connected to the troubleshooting problem. This simulation consists of a *modem, computers, telephone, fax, adsl filters,* etc. In addition, parts of these objects that will be of interest in the troubleshooting process have also been simulated. For instance, the *powerlight, adsl light on the modem, the usb ports on the computer, modem software,* etc are represented. Since connections between these objects cause problems in the real world, they are also simulated. In addition to representing the objects in the environment, the simulation allows the user to access the objects. Just like in the real world, users are able to manipulate the objects and therefore change the state of the environment as a whole. For instance, rebooting the modem might set the internet connection correctly. In the current model, the user will be able to observe and manipulate the objects in the environment in a principled way. In real world troubleshooting practice, the experts will be able to ping the user's modem remotely. This calls for a manipulative interface between the expert and the environment. This feature is not available in the current environment model.

The environment simulation is represented using Prolog facts and rules. The state of the environment S_e is represented using dynamic facts and is set initially using Environment-setting rules. The state of the environment can be observed using observation rules and the environment can be manipulated using manipulation rules. Example rules are shown in table 4.1. The environment-setting rule shown sets the environment initially to one of the faulty scenarios (faulty phone filter) that the current model is able to simulate. The *adsl filter a2* that connects the *phone_socket* to the *telephone t1* is specified as faulty in the definition. This causes interference and is the source of the problem. Using observation actions, the user is able to observe that the *adsl filter* is present. As per the system's instructions, the user can use manipulative actions to fix this problem. In this case, he replaces the *phone adsl filter* and sets the connection correctly.

Table 4.1 Environment Simulation

Environment Setting rule

```
set_env1(faulty_phone_filter) :-
        assert(equipment(comp1, desktop, working)),
        assert(equipment(t1, phone, working)),
        assert(equipment(m1, modem, working)),
        assert(equipment(a1, adsl_filter, working)),
        assert(equipment(a2, adsl_filter, not_working)),
        assert(connected(phone_socket, a1, rj11, firm)),
        assert(connected(phone_socket, a2, rj11, firm)),
        assert(connected(a1, m1, rj11, firm)),
        assert(connected(a2, t1, rj11, firm)),
        assert(connected(m1, comp1, usb, working)),
        assert(phone_line(live)),
        assert(modem_software(installed, working)),
        assert(authentication(correct)),
        !.
```

Observe Environment : Is there a phone filter?

```
pact(adsl_filter_for_phone, present) :-
        equipment(P, phone, _),
        equipment(A, adsl_filter, _),
        connected(A, P, rj11, firm),
        !.
```

Manipulate Environment : Replace the phone filter

```
mact(replace_phone_filter) :-
        equipment(a2, adsl_filter, not_working),
        !,
        retract(connected(phone_socket, a2, rj11, firm)),
        retract(connected(a2, t1, rj11, _)),
        assert(equipment(a4, adsl_filter, working)),
        assert(connected(phone_socket, a4, rj11, firm)),
        assert(connected(a4, t1, rj11, firm)),
        update_env,
        !.
```

The current environment simulation is capable of simulating 6 error configurations - *faulty modem, faulty phone/modem filter, missing phone/modem filter, faulty phone line, authentication failure, faulty modem USB port*.

5 User Simulation

The user simulation model stochastically simulates environment-sensitive and concept-knowledge-

sensitive user behavior, as shown schematically in figure 5.1 and described below.

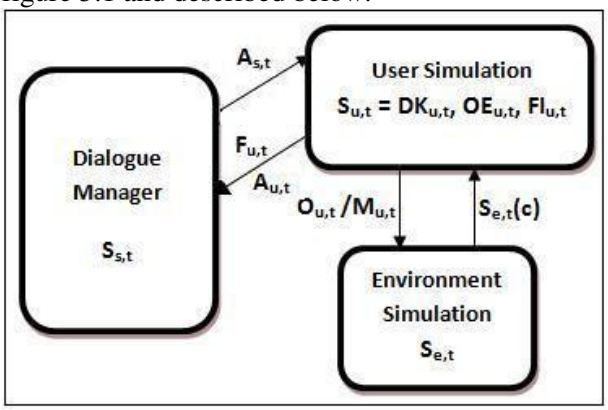


Figure 5.1. The Model & Experimental setup

Besides being faithful to the environment, it also simulates the *domain concept knowledge* of the user DK_u. Therefore, when the user is confronted with concepts he barely knows, he is more likely to request clarification. The model is capable of updating its domain knowledge, when the system provides clarification. By initially selecting different knowledge profiles, the current simulation can simulate a continuum of users from novices to experts. We attach a probability to every concept that the users must know in order to co-operate with the system to fix the problem. This probability determines whether the user follows the system's instruction or requests clarification. The knowledge profiles of three types of users are given in table 5.1. These values would be set by an analysis of data collected for the specific troubleshooting domain – for this proof of concept we select illustrative values.

Table 5.1. User knowledge profiles: P(DK)

Concept\User	Expert	Intermediate	Novice
Phone_line	0.99	0.85	0.5
ADSL filter	0.8	0.6	0.2
ADSL light	0.9	0.6	0.2
USB slot	0.8	0.55	0.3
Powerlight	0.95	0.85	0.5

An expert user has the highest probability to answer queries and perform manipulative actions without requesting clarification. However these values are not static profile thresholds but initial values during the start of a dialogue session. These values increase when the system clarifies the concept by issuing an extended request. In this case, for the particular concept *c* under discussion:

$$P(DK_{u,c,t+1}) = P(DK_{u,c,t} + boost\_c)$$

For these experiments we set *boost_c* = 0.5 for each c. Again, this average probability boost could be determined from real data for each c. In an extended request the system spends a few utterances to educate the user and therefore update the user's domain knowledge. As their knowledge profile gets boosted, their chances of answering the query increase. For instance, in response to the DM's action *'request_info: adsl_light'*, a novice user is more likely to say *'request_clarification: adsl_light',* initially. But if the system clarifies the concept, they are more likely to say *'provide_info: on'*. However an expert user is more likely to return *'provide_info: on'* without requesting clarification (given that the ADSL light is on as per the environment state).

Table 5.2 User's Action set A_u

1. Provide_info
2. Acknowledge
3. Manipulate_and_acknowledge
4. Request_clarification
5. Hang_up

When the system requests information, the user simulation observes the environment by issuing an observation action O_u, updates its observations OE_u, and reports back to the system. Similarly, when the system requests that the user manipulate the environment, the user manipulates the environment using a manipulate action M_u, observes its effects, updates the observations OE_u, and reports it back to the system. A manipulate action also changes the state of the environment S_e. The interaction between the user, environment and the system is shown in fig 5.1.

In addition to the user's verbal response, we also simulate the user's frustration. A user gets frustrated when the dialogue manager does not explain unknown terms or unnecessarily explains well- known terms[1]. We use a frustration index FI_u to capture this behavior. This index affects the probability of the user hanging up the call before the dialogue ends. The probability of hang-up is twice their frustration index, as a probability. The

[1] However, weighing of under-informativeness and over-informativeness on the same scale may be revised in our future work.

user's frustration is also conveyed to the system on a turn by turn basis $F_{u,t}$ (as a boolean value) along with the user's action $A_{u,t}$. We assume that the dialogue manager can detect the user's frustration from the user's utterance. It has been shown that frustration can be determined from prosodic (Lee et al 2002) features in the user's utterance.

The user state S_u, contributing to the action selection process contains the user's domain knowledge DK_u, his observations of the environment OE_u, frustration index FI_u and the number of dialogue turns T. Each of these components is updated at different times during the action selection process.

$$S_u = <DK_u , OE_u , FI_u , T>$$

Based on the updated user and environment states and the system action, the user action A_u is selected as described in the following algorithm, where *return_action P(A|X)* returns a user action A with probability P(A|X):

Step 1: If turns T > 6,
return_action P(hang_up | FI_u)
Step 2: If $A_{s,t}$ = extended_request(c),
Update $DK_{u,c}$, FI_u
Step 3: If $A_{s,t}$ = request or extended_request(c),
Do P(O_u(c) | DK(c)) or P(M_u(c) | DK(c))[2]
Update S_e, $OE_{u,c}$
return_action P($A_{u,t}$ | OE_u)

Figure 5.3 User simulation algorithm

Thus, by conditioning the user's action A_u on various factors like the user's domain knowledge DK_u, user's observations OE_u, frustration FI_u and the environment state S_e, the simulation provides a context-consistent and diverse user behavior. In future, we also plan to validate the simulation against real user datasets using well established metrics (Schatzmann et al 2005).

6 Reward Function

Every dialogue session is rewarded at its completion. A task completion reward (TCR) of 500 is given for successful completion and -500 for unsuccessful dialogues. A dialogue is considered to be successful if the dialogue partners are able to fix the problem and close the dialogue. On the other hand, it is considered unsuccessful if the user gets frustrated and hangs up before the dialogue ends. The following costs are associated with the number of dialogue turns (T) and extended turns (ET).

Turn cost per turn: TC = 10.0
Extended cost per turn: EC = 30.0

This extended turn cost encodes that idea that extended turns cost 3 times as much as normal turns, on average. Again, this parameter would be set by a PARADISE-style (Walker et al 1997) regression analysis on real user data.

The final reward for a dialogue session is calculated based on:

Total Turn cost: TTC = T * TC
Total Extension cost: TEC = ET * EC
Final reward = TCR - TTC – TNC

The reward function is designed to penalize longer dialogues and unnecessary extensions. The task of the learning agent (dialogue manager) is to learn an optimal policy to minimize the costs and increase the chances of successful dialogue of for all kinds of users.

7 Training

The system was trained for 15000 cycles producing approximately 1500 dialogues using the SARSA reinforcement learning algorithm (Sutton & Barto 1998). Our objective was to learn a single policy that can adapt online to any type of user (novice, expert, or intermediate). Hence the user simulation was calibrated to produce an equal number of novice and expert users. Recall that these types of user behave stochastically, so that no two expert users are guaranteed to behave in the same way, for example. The users were allowed to hang up only after the sixth turn. This is manually chosen to avoid early hang ups. After the sixth turn, the probability of the user hanging up is directly proportional to the user's frustration index (as described above). The main learning task here is to decide between using *extended- request* and *request* acts.

After the training runs, the system learned to adapt online to both expert and novice users and in each case to maximize the final reward. It learned to effectively make use of the user expertise index, which is a part of the dialogue state, in order to tune its behavior towards the users. It learned to use *extended-request* acts when the expertise index indicates that the user is a novice and to use *request* acts when it indicates an expert user. It also learns not to use *extended-request* acts for information

[2] Depending whether the system has requested an observe or manipulate action.

that does not depend on the user's conceptual knowledge. For instance, both novice and expert users are equally able to answer if their internet connection is working. Hence an *extended-request* act simply adds to the cost without reducing any risks or costs and so is avoided for such slots. The system learned to reduce the dialogue length during training (shown in fig 7.1) by choosing the appropriate action the very first time an instruction is given to a user. This behavior avoids repeating the instruction and therefore the costly repairs.

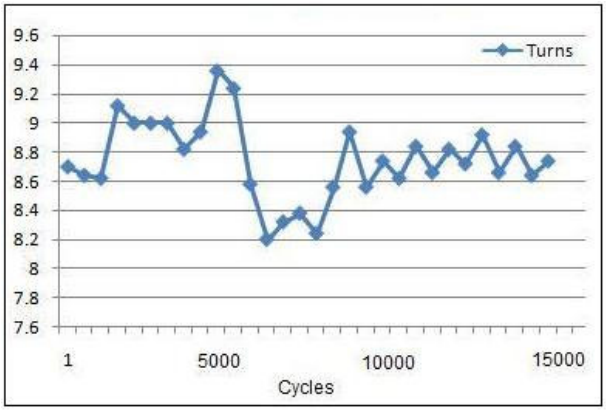


Figure 7.1. Optimization of dialogue length

Similarly, during training, the dialogue manager learned to optimize the user frustration index (shown in fig 7.2).

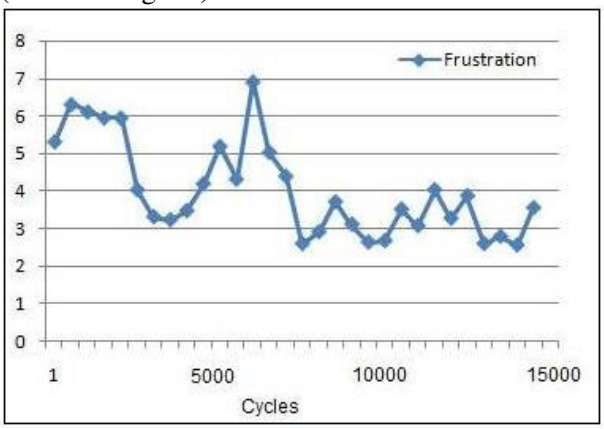


Figure 7.2. Optimization of frustration index

8 Evaluation

We tested the learned policy by comparing its performance with three hand-coded policies - Expert Only, Novice Only, and Adaptive.

1. The *Hand-coded Expert Only* policy treats all users as experts. Hence it always uses request acts.
2. The *Hand-coded Novice Only* policy treats all users as novices and always uses extended-request acts.
3. The *Hand-coded Adaptive* policy adapts to the user based on the estimated expertise index. It uses *extended-request* acts when the index is less than 5 and *request* acts otherwise. This therefore encodes a classic threshold-based approach to adaptation. (see e.g. Varges 2003)

All four policies were run against three groups of users – experts, novices and intermediate users. Each such run produced approximately 800 test dialogues each. Task success rate and average final reward for each policy run on different user groups were calculated. The results for an equal mix of experts, novices, and intermediate users are presented under "Mixed". Statistical significance was calculated using a Wilcoxon signed rank test[3].

Table 8.1 compares the task success rates of different policies on the user populations. Note that the most important column in the following tables is "Mixed", since this indicates the performance of the policies when they encounter a mixed population of users (i.e. as would happen in a real deployed system).

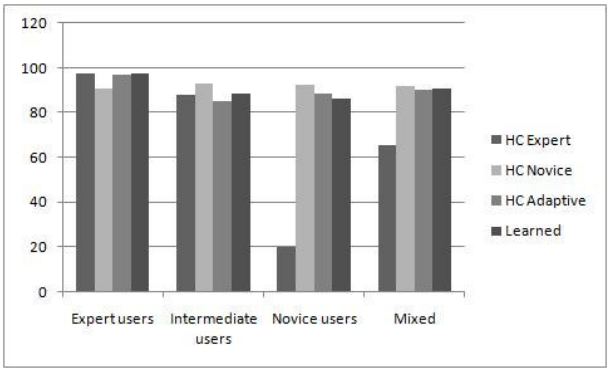


Figure 8.1 Task success rate

Table 8.1. Task success rate

Policy\Users	Expert	Inter	Novice	**Mixed**
HC Expert	**97.7**	87.9	19.5	65.1
HC Novice	90.8	**92.8**	**92.6**	**92.1**
HC Adaptive	96.9	85.2	88.3	90.1
Learned	**97.6**	**88.2**	**86.2**	**90.6**

[3] Difference in average final rewards between the policies were not normally distributed as per Kolmogonrov-Smirnov test.

Here we can see that the learned policy has good task completion across the range of user types. However, the task success metric does not take into account the important cost of different types of system turns. Table 8.2 compares the average final reward (combing task completion and turn cost) of the different policies on three different user populations. All improvements made by the learned policy are statistically significant at $p<0.001$.

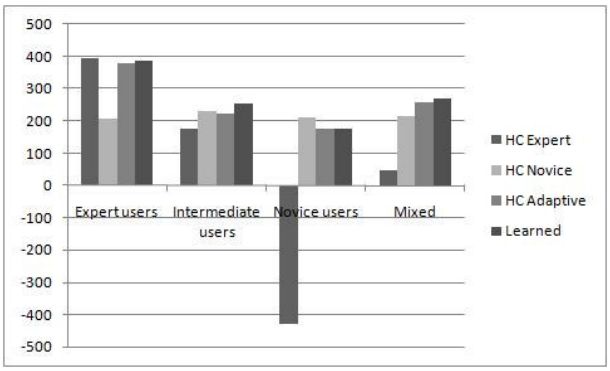


Figure 8.2 Average final reward

Table 8.2: Average final reward				
Policy\Users	Expert	Inter	Novice	**Mixed**
HC Expert	**393.5**	177.5	-430.2	46.9
HC Novice	206.6	230.7	**210.1**	215.8
HC Adaptive	379.6	221.7	177.2	259.4
Learned	**385.9**	**252.6**	**175.2**	**271.2**

The HC Expert Only and HC Novice Only policies were best for their respective populations, but they did not perform well against other populations. The HC Expert Only policy beats all the other policies with the highest average final reward 393.5 for the expert users. It also has the highest task success rate at 97.7, but it performs badly on novice and intermediate users, scoring the lowest average final reward among other policies. The HC Novice policy performed well with both novice and intermediate users, but with expert users it scored the lowest average final reward among all the policies. It also has the highest task success rate for all users combined. This is because it always gives the user an extended request, thereby reducing the number of turns. But this also results in a reduction in average final reward on all user types combined (Mixed). The HC Adaptive policy performs consistently among all the user groups. It scores a very good average final reward and task success rate. But it scores lower than the learned policy in both average final reward and task completion scores in expert and intermediate user groups. However, in the novice user group, its scores are slightly better than the learned policy. The learned policy also performs consistently on all the groups. It scores the best average final reward for intermediate users, although the policy was not trained on intermediate users. Its average final rewards on novice and expert groups are not very far behind the best rewards. However, the key point here is that for the "mixed" user group, the learned policy beats all the other policies with the highest average final reward of 271.2. This result shows that when we do not know the user population in advance, as is the case in real applications, the learned policy is able to handle the range of users encountered by adapting online. This is consistent with the results of Lemon & Liu (2007), who considered dialogue policies for different noise conditions.

The above results are promising because the learned policy has been able to perform better than carefully handcrafted adaptive policy for the same task. While it was easy to hand-code a policy for this task, it would not be so when more parameters are added to the dialogue manager's information state. The policy learning paradigm allows us to learn optimal policies for these types of trade-off without an expensive "implement, test, deploy, refine, redeploy,..." iterative development cycle. The parameters for the model presented above can all be estimated from data, for example collected in a small Wizard-of-Oz experiment (Rieser and Lemon 2008).

9 Conclusion

We addressed the general problem that different dialogue participants have different types of conceptual knowledge of the domain under discussion, so that work must be done to align or coordinate their understanding (see e.g. Issacs & Clark 1987). We studied this problem in the practical case of "troubleshooting", where automated dialogue systems need to adapt online to the different knowledge of different users as it encounters them. We proposed a new user model which incorporates different conceptual knowledge of different users, together with an environment simulation. We show that this model allows us to learn dialogue policies that automatically adapt online to new users, and that these policies are significantly better than threshold-based adaptive hand-coded policies for this problem. Future work in this area would be to ex-

tend the approach, and in particular increase the complexity of the user simulations, to handle other aspects of alignment in dialogue, such as semantic plasticity (Larsson 2007) and lexical and syntactic alignment (Pickering and Garrod 2006) in task-based dialogues.

10 Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework (FP7/2007-2013) under grant agreement no. 216594. (CLASSIC Project www.classic-project.org), EPSRC project no. EP/E019501/1 and the British Council (UKIERI Ph.D Scholarships 2007-08).

References

H. Cuayáhuitl, S. Renals, O. Lemon, and H. Shimodaira. 2005. Human-Computer Dialogue Simulation Using Hidden Markov Models. *Proc. of ASRU* '05.

W. Eckert, E. Levin and R. Pieraccini. 1997. User Modelling for Spoken Dialogue System Evaluation. *Proc. of ASRU 1997.*

K. Georgila, J. Henderson, and O. Lemon. 2005. Learning user simulations for Information State Update Dialogue Systems. *Proc. Eurospeech-Interspeech '05*.

J. Boye. 2007. Dialogue management for automatic troubleshooting and other problem-solving applications. *Proc. 8th SIGDial workshop on discourse and dialogue*, 2007.

E. A. Isaacs & H. H. Clark. 1987. References in conversations between experts and novices. *Journal of Experimental Psychology: General, 116, 26-37.*

S. Larsson. 2007. A general framework for semantic plasticity and negotiation. *Proceedings of the 7th Intl Workshop on Computational Semantics.*

C. M. Lee, S. Narayanan, R. Pieraccini. 2002. Classifying emotions in human-machine spoken dialogs. *Proc. of ICME*, 2002.

O. Lemon, X. Liu. 2007. Dialogue Policy Learning for combinations of Noise and User Simulations: transfer results. *Proc. 8th SIGdial Workshop, 2007*.

O. Lemon. 2008. Adaptive Natural Language Generation in Dialogue using Reinforcement Learning. SEMDial (LONDial) 2008.

M. J. Pickering and S. Garrod. 2006. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27, 169-225.

O. Pietquin and T. Dutoit. 2006. A probabilistic framework for dialog simulation and optimal strategy learning. *In IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 589–599, March 2006.

V. Rieser and O. Lemon. 2008. Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz data: Bootstrapping and Evaluation, *In ACL 2008* (to appear).

J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young. 2007. Agenda-based User Simulation for Bootstrapping a POMDP Dialogue System, *Proc. HLT/NAACL, 2007.*

J. Schatzmann, K. Georgila, and S. Young. 2005. Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems. *Proc. 6th SIGDial Workshop on Discourse and Dialogue, Lisbon, 2005.*

K. Scheffler & S. Young. 2001. Corpus-based dialogue simulation for automatic strategy learning and evaluation. *Proc. NAACL Workshop on Adaptation in Dialogue Systems, 2001.*

R.S. Sutton, A.G. Barto. Reinforcement Learning : An Introduction. *MIT Press, 1998.*

S. Varges. 2003. Instance-based Natural Language Generation. *Ph. D. Thesis. Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.*

M. Walker, D. Litman, C. Kamm, and A. Abella. 1997. PARADISE: A Framework for evaluating Spoken Dialogue Agents. *Proc 35th Annual meeting of ACL.*

J. Williams. 2007. Applying POMDPs to Dialog Systems in the Troubleshooting Domain. *Proc HLT/NAACL Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology* 2007.

Following Assembly Plans in Cooperative, Task-Based Human-Robot Dialogue

Mary Ellen Foster
Informatik VI: Robotics and Embedded Systems
Technische Universität München
foster@in.tum.de

Colin Matheson
School of Informatics
University of Edinburgh
colin@inf.ed.ac.uk

Abstract

The JAST dialogue system allows a human and a robot to jointly assemble construction toys on a common work area. Supporting this type of dialogue requires that the system have a representation of assembly plans that permits it both to discuss the details of the plan and to monitor its execution. We present a conceptual representation of assembly plans based on AND/OR graphs, and then describe how the dialogue manager uses these plans as the basis for a range of strategies for jointly carrying out the plans with the user.

1 Introduction

An increasing number of interactive systems are addressing the task of supporting intelligent cooperation with a human partner, where both partners work together to achieve a mutual task. This type of task-based collaboration is particularly relevant for robots, which are able to sense and perform actions in the physical world and can often be treated as full team members (Breazeal et al., 2004; Fong et al., 2005). For an artificial system to be able to work together with a human on such a task, the details of the task must be represented in such a way that the system can both follow the task progress and participate in discussing the details of the task execution.

In this paper, we present the JAST human-robot dialogue system, which allows the user to cooperate with the robot in assembling wooden construction toys. Assembly plans are represented as AND/OR graphs (Homem de Mello and Sanderson, 1990), which is the standard mechanism for representing such plans in autonomous robot assembly. This representation allows the dialogue manager to access the current steps in the plan and to update the state of the world following user actions. The dialogue manager implements two strategies for explaining a plan to the user, one that traverses the plan in a depth-first way, naming objects after they are complete, and another that names and describes the objects top-down.

The interactions supported by the JAST system is quite similar to the 'Max' virtual communicator system developed at the University of Bielefeld (Kopp et al., 2003; Rickheit and Wachsmuth, 2006). However, the mechanisms underlying the interactions are different: while the core of Max is a cognitively-motivated agent architecture, JAST uses a dialogue manager based on the information-state update paradigm. Our implementation also shares some features with Blaylock and Allen (2005)'s *collaborative problem-solving* (CPS) model of dialogue. That model divides the problem-solving process into three general phases: determining objectives, determining and instantiating recipes, and executing recipes and monitoring success. While we do not employ the full formal structure of the CPS model, the JAST system views collaborative dialogue in a similar way. A similar link between domain plans and dialogue strategies is also used in the LeActiveMath mathematics tutorial dialogue system (Callaway et al., 2006) to allow the system to describe and cooperatively follow plans drawn from a domain reasoner and to give context-specific hints to guide a learner through the graph of a solution.

2 The JAST human-robot dialogue system

The overall goal of the JAST project ('**J**oint **A**ction **S**cience and **T**echnology') is to investigate the cognitive and communicative aspects of jointly-acting

Figure 1: The JAST dialogue robot

agents, both human and artificial. The JAST human-robot dialogue system (Rickert et al., 2007) is designed as a platform to integrate the project's empirical findings on cognition and dialogue with its work on autonomous robots, by supporting multimodal human-robot collaboration on a joint construction task. The user and the robot jointly assemble wooden construction toys on a common workspace, coordinating their actions through speech, gestures, and facial displays.

The robot (Figure 1) consists of a pair of mechanical arms with grippers, mounted in a position to resemble human arms, and an animatronic talking head able to produce facial expressions, rigid head motion, and lip-synchronised synthesised speech. The input channels consist of speech recognition, object recognition, robot sensors, and face tracking; the outputs include synthesised speech, head motions, and manipulator actions.

In the current version of the system, the robot is able to manipulate objects in the workspace and to perform simple assembly tasks. The primary form of interaction with the current system is one in which the robot instructs the user on building a particular compound object, explaining the necessary assembly steps and retrieving pieces as required; at the end of the paper, we discuss extensions to this scenario. To make joint action essential to the assembly task, the workspace is divided into two areas: one belonging to the robot and one to the human. The pieces necessary for building a desired assembly are distributed across these areas so that neither of the agents is able to reach all of the required components and must rely on the partner to retrieve them.

3 Representing assembly plans

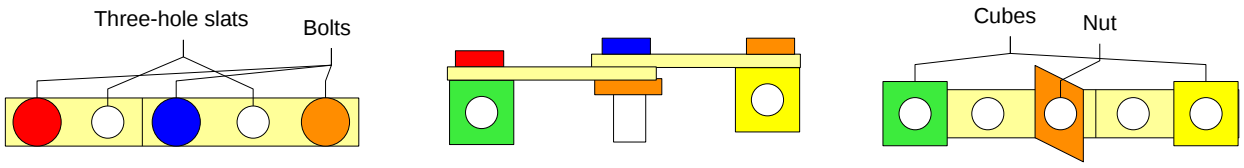
Like several previous interactive systems designed to support (physical or virtual) joint assembly (e.g., Knoll, 2003; Rickheit and Wachsmuth, 2006), dialogues in JAST are based around assembling *Baufix* wooden construction toys. The following are the basic components that are available:

- Threaded **Bolts** of varying lengths and colours;
- **Cubes** of varying colours, with four threaded holes and two unthreaded holes;
- **Nuts** with a single threaded hole; and
- **Slats** with three, five, or seven unthreaded holes

For the remainder of this paper, we will consider the sample object shown in Figure 2, which we will call a 'bridge'. This object consists of two small (three-hole) slats, connected end-to-end using a blue bolt and a nut, with a cube connected to the other end of each slat. Some of the sub-components also have names: the slat+cube combination on the left of Figure 2 is called the 'front', while the combination on the right is called the 'back'.

Even for this fairly simple object, there are a number of different possible assembly sequences: the slats may be joined together at any point, and the two cubes can also be attached in any order. There is some symmetry in the plan: for example, the two slats are interchangeable, and it is not important which end of a slat or which hole in a cube is used. However, there are also geometric relationships among the pieces that must be respected, such as the fact that the bolts all go through the slats in the same direction.

In this section, we present the assembly-plan representation used in JAST, which captures all of these features. We begin by describing the representation of individual assembly steps and then show how those steps are combined to describe the full plan.



(a) Top view (b) Side view (c) Bottom view

Figure 2: Assembled object ('bridge')

3.1 Assembly steps

The first step in representing an assembly plan is to represent the individual assembly steps. In our system, assembly steps are represented in a domain-specific way, tailored to the types of objects that can be constructed from Baufix components. Following Sagerer et al. (2002)—who also represented Baufix assemblies for use in interactive assembly—we define an assembly step to consist of the following components:

- Exactly one bolt;
- Any number of unthreaded pieces to insert the bolt through (cubes, slats, or composite objects containing cubes or slats); and
- Exactly one threaded fastener to screw onto the bolt (a nut, a cube, or a composite object containing a cube).

In addition to the above features, an assembly-step description also includes details to ensure that the step is performed correctly. These details fall into three main classes:

- For any component that is a composite object, which piece of that object should be used;
- Which of the several holes in a slat should be used; and
- The direction of insertion or fastening.

Note that not all of these details are necessary for a single step: it does not matter which of the four threaded holes in a cube is used for an attachment operation, for instance, and the two end holes of a slat are also interchangeable, as are the two faces. However, when the same component is used in more than one assembly step—as in the sample object,

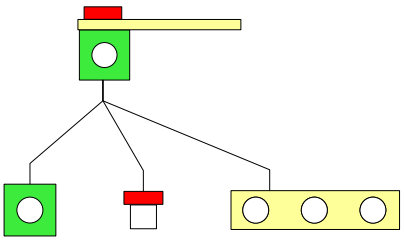


Figure 3: Single assembly step

Bolt $b_1(\text{bolt}, \text{small}, \text{red})$
Insert list $[i_1(\text{slat}, \text{three-hole})]$
Fastener $f_1(\text{cube}, \text{green})$
Details $[\text{hole}(i_1) = \textbf{Middle}, \text{direction}(f_1) = \textbf{South}, \text{direction}(i_1) = \textbf{South}]$

Figure 4: Symbolic description of the assembly step

where each of the two slats is used twice—it is important that all of those steps are performed based on the same frame of reference to ensure that the relative positions of the objects are correct. We therefore define a canonical orientation of an assembled object (as in Figure 2(b)).

Figure 3 illustrates the assembly step that creates the 'front' of the sample object: a red bolt is inserted from above through the end of a three-hole slat and is then screwed into a threaded hole of a green cube. Figure 4 gives a symbolic description of this step.

3.2 Assembly plans

Each possible assembly plan for an object is made up of a sequence of assembly steps, where a single object may have a number of such sequences. In autonomous assembly, the standard solution for representing such a set of assembly sequences is the *AND/OR graph*: a directed acyclic graph that decomposes a problem into two sets of nodes, AND nodes and OR nodes. An AND node is satisfied

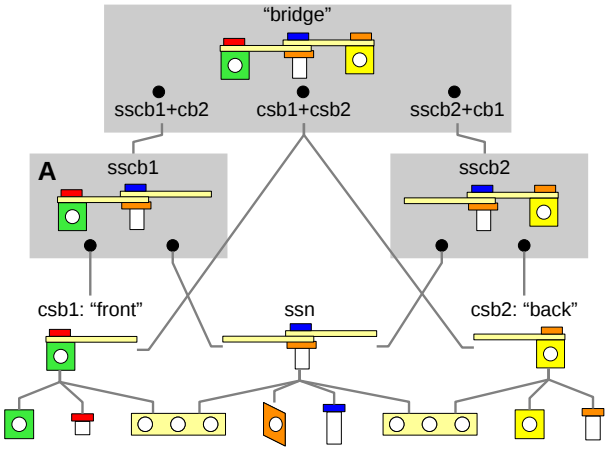


Figure 5: AND/OR graph structure for the sample object

only if all of its children are satisfied, while an OR node is satisfied when exactly one of its children is. This provides a natural representation for any problem that can be represented by decomposing a goal into subgoals, and was first proposed for robot assembly by Homem de Mello and Sanderson (1990).

In an assembly plan, an AND node corresponds to a single assembly step in which all of the children are combined to produce a more complex component. An OR node, on the other hand, corresponds to situations in which an assembly may be produced by different sequences of assembly operations; in this case, each child of the node corresponds to a different assembly sequence. An AND/OR graph provides a compact representation of all of the possible assembly sequences for an object; each individual sequence can be extracted by traversing the tree top-down, including all of the children of each AND node and exactly one child of each OR node.

Figure 5 shows the structure of the AND/OR graph for the sample object in Figure 2. Nodes with outgoing edges represent AND nodes, while nodes highlighted with a grey background are OR nodes. The leaf nodes in the tree correspond to the individual pieces required to build the sample object, while each internal node corresponds to a subassembly. For example, the subtree rooted at the OR node marked **A** indicates that there are two different assembly sequences that can result in that component. The first, corresponding to the left child of **A**, involves first attaching the green cube to one end of the slat to make the 'front' and then attaching the other slat to the other end. The second sequence, corresponding to the right child of **A**, first creates the (unnamed) centre piece and then attaches the cube to the end.

Each internal node has a unique ID—for example, the node corresponding to the assembly operation from Figure 3 has ID *csb1*. Three of the nodes also have labels indicating that the corresponding sub-assembly has a name: the 'bridge', the 'front' and the 'back'.

Previous systems have also addressed the task of representing assemblies of Baufix-style objects. Brock (1993) represented components by their geometric properties and described assemblies in terms of hierarchical planning operators. This system had the goal of creating plans for a robot to autonomously assemble the components, with no user interaction. Sagerer et al. (2002) used a similar representation for assembly actions to the one described here, with the goal of recognising complex objects in an interactive human-robot scenario. This system did not represent full assembly plans, but rather structural descriptions sufficient for recognition. The representation for Baufix assemblies in for the Max virtual communicator Jung (2003) described them in terms of *ports* and *connections* of CAD-based parts with the goal of supporting assembly in virtual environments.

The JAST representation described in this section is most similar to that used by Sagerer et al. (2002), although they do not use AND/OR graphs to represent the assembly plans; the other representations concentrate more on detailed geometric features that are less relevant to the current scenario where the user is the primary agent for assembly operations.

4 Following an assembly plan in dialogue

Interactions in JAST are based around cooperatively carrying out assembly plans represented as described in the preceding section. As mentioned earlier, in the current scenario, the robot is aware of the target object and the full plan and instructs the user on carrying out the assembly, and the user learns to make particular sub-components along the way. In Section 5, we discuss possible extended interactions, but in this section we concentrate on the robot-as-instructor scenario. Excerpts from typical interac-

Depth-first	
SYSTEM[1]:	First we need to build a bridge. Okay?
USER[1]:	Okay
SYSTEM[2]:	[*picking up green cube*] Insert the red bolt into the end of a slat and fasten it with this cube.
USER[2]:	Okay
SYSTEM[3]:	Well done. You have completed the front. Now insert ....
Top-down	
SYSTEM[1]:	First we need to build a bridge. Okay?
USER[1]:	Okay
SYSTEM[2]:	To build a bridge we need to make a front and a back. To make a front, insert the red bolt ...

Figure 6: Example depth-first and top-down interactions

tions using two different explanation strategies are shown in Figure 6. In the remainder of this section, we describe how the components of the system work together to support such interactions.

The required knowledge is distributed across three main components of the system. The **task planner** stores and maintains the AND/OR graph corresponding to the current plan, updating it as appropriate based on information from other modules. The **object inventory** tracks the properties and locations of Baufix objects in the world, using information from the object-recognition system as well as the task planner. Finally, the **dialogue manager** (DM) receives a unified representation of user speech and actions from the input-processing components and selects appropriate system output based on the current state of the interaction and of the plan, along with the user's assumed knowledge. It also updates the state of other components based on the events in the dialogue.

The DM is based on the TrindiKit dialogue-management toolkit, which uses the information-state update approach to dialogue management (Traum and Larsson, 2003). The JAST information state (IS) includes data about the user's knowledge, the current step in the plan that is being executed, the history of steps that have been described to the user, and the history of the interaction. Figure 7 in Section 4.2 below contains an example IS and some further discussion.

Table 1: Initial object inventory

ID	Type	Properties	Location
1	Bolt	Color=Red	Table(User)
2	Bolt	Color=Orange	Table(Robot)
3	Bolt	Color=Blue	Table(Robot)
4	Cube	Color=Yellow	Table(Robot)
5	Cube	Color=Green	Table(Robot)
6	Cube	Color=Green	Table(User)
7	Nut	Color=Orange	Table(Robot)
8	Slat	Size=3-hole	Table(User)
9	Slat	Size=3-hole	Table(User)
10	Slat	Size=5-hole	Table(User)

4.1 Loading the plan

Before the first utterance in the dialogue excerpt, the system must select a target object to assemble. In our scenario, where the robot knows the plan and must instruct the user, the choice of target object is fixed, so the DM simply instructs the task planner to load the AND/OR graph for the target object from its library of fully-specified plans. At the moment, the task planner also selects a specific assembly sequence from the AND/OR graph at the point that the plan is loaded, favouring sequences that include more named sub-components (e.g., the 'front') to ensure that the user learns to build them.

4.2 Describing assembly steps to the user

Once the AND/OR graph has been loaded and a sequence selected, the DM must describe the assembly process to the user. The DM can proceed **depth-first**, describing plan steps and naming objects when they are complete, or it can work **top-down** and describe and name each step in advance; both of these strategies are illustrated in Figure 6. In both cases the actual path through the plan is the same, and the 'current state' of the dialogue as represented in the IS is a crucial component. A truncated example of an IS (relevant to either strategy) is contained in Figure 7 and shows the basic plan information and typical dialogue history (DH) and user knowledge (UK) representations. The names of the plan nodes (e.g., 'front' or 'bridge') are associated inside the planner with (language independent) concepts which the language generation system turns into lexical items in English or German; the DM only needs to know the plan node identifier. The DH contains an ordered

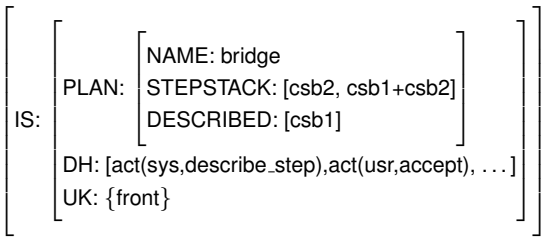


Figure 7: Part of the IS Structure

list of completed acts, and the user knowledge is represented as a set of 'known' object types. This set is maintained throughout the dialogue, so if we are constructing an object for the second time we can ask the user if they remember how to build it.

In operation, the DM first requests the children of the tree node corresponding to the step to be described. A check is carried out to determine whether all the objects mentioned in the step are either basic components (bolts, cubes) or known to the user; if not, the system picks the left corner child and iterates until such a node is found. If the depth-first strategy is being pursued, the DM proceeds without producing linguistic output until it reaches a node where everything necessary to build the object exists. In the top-down approach the system names each node and describes the general structure as it proceeds, whereas the depth-first strategy assumes that objects are built before they are named. The 'delayed naming' aspect of the depth-first approach is not, of course, necessary; it is perfectly possible to tell the user what is being constructed before it is described. However, the naming strategy is an aspect of dialogue that we would like to experiment with, and it seems less natural to combine top-down description with delayed naming.[1] As each step is completed, the DM sends the planner a 'step executed' message with the relevant node name, which updates the system state as described in Section 4.5. When the whole plan is complete, the system loads the next assembly plan or terminates.

[1]Top-down with delayed naming would suggest system utterances such as:

> Let's build a bridge. Insert a blue bolt through a green slat and fasten it with a yellow cube. This is a front. Now insert ...

From the perspective of Centering Theory (Grosz et al., 1995), the focus shifts are non-optimal.

```
<rst>
  <consequence id="id1">
   <item idref="id2" />
   <item idref="id3" />
  </consequence>
  <item id="id2" type="impersonal">
   <pred action="build" result="front" />
  </item>
  <join id="id3">
   <item idref="id4" />
   <item idref="id5" />
  </join>
  <item id="id4" type="imperative">
   <pred action="insert">
           .... contents of insert ....
   </pred>
  </item>
  <item id="id5" type="imperative">
   <pred action="fasten">
           .... contents of fasten ....
   </pred>
  </item>
 </rst>
```

Figure 8: An RST 'Consequence' Structure in XML

4.3 System Output

The DM builds XML structures containing RST-style representations (Mann and Thompson, 1988) to be passed on to the output planner and ultimately the language generator. The top-down strategy uses 'consequence' relations to link the actions being described and their results, as illustrated in Figure 8. The 'insert' and 'fasten' elements in the figure, which describe the objects, have been removed for brevity. The type attribute on item elements specifies the basic clause class; impersonal clauses such as 'to build a bridge', imperatives such as 'insert the bolt', and declaratives as in 'the bridge is complete'.

An important aspect of describing a step to the user is selecting an appropriate means of referring to the required objects, which is performed by the output planner and depends on the information in the object inventory. The initial object inventory for the sample interaction is shown in Table 1. When the system generates the SYSTEM[2] utterance in the second (top-down) extract, the back and the front do not yet exist, so they are referred to indefinitely. However, there is one red bolt on the user's table, so a definite is used, while there are 3 relevant slats, so again an indefinite is appropriate. The robot has selected a cube to pick up (if more than one available object matches the description the choice is random), so in this case a demonstrative is used.

4.4 Responding to user actions

Currently the user may respond in a restricted number ways to a system utterance; the range will be extended in the near future, but for now we allow verbal acknowledgements of various kinds, indications of misunderstandings, and yes-no answers. The sample dialogues in Figure 6 contain examples of acknowledgements which are interpreted in different ways. Following SYSTEM[1], 'okay' is interpreted mainly as indicating understanding, while following SYSTEM[2] 'okay' is assumed to indicate that the user has performed the actions described. Depending on the confirmation strategy used by the system, such interpretations might be queried explicitly; the balance between verbal confirmation and the current optimistic grounding approach is another area for experimentation.

The user can indicate that something is misunderstood, in which case previous output is typically repeated. The user may also be asked yes-no questions, which are again interpreted differently depending on the dialogue context. The most obvious example is in cases where the DM reaches a plan step whose result is already listed in the user knowledge set. In this situation the system has the option of asking the user whether or not they remember how to build the object in question.

4.5 Updating the state

Once an assembly step has been completed, the state of the task planner must be updated. As noted above, the DM informs the rest of the system that the step has been executed. This message includes the IDs of the objects that were used, and in response the task planner performs two actions: it updates the set of world objects in the inventory, and it marks the step as completed in its internal AND/OR graph.

Completing an assembly step has two effects on the object inventory. First, all of the components that were involved in the assembly are no longer available for use, so their location is adjusted to indicate that they are part of a larger component. Second, a new object is introduced into the world corresponding to the sub-assembly that was created by the completed step. The updated object inventory after USER[2] in Figure 6 is shown in Table 2, with objects changed by the action indicated by italics.

Table 2: Updated object inventory

ID	Type	Properties	Location
1	*Bolt*	*Color=Red*	*Assembled(11)*
2	Bolt	Color=Orange	Table(Robot)
3	Bolt	Color=Blue	Table(Robot)
4	Cube	Color=Yellow	Table(Robot)
5	*Cube*	*Color=Green*	*Assembled(11)*
6	Cube	Color=Green	Table(User)
7	Nut	Color=Orange	Table(Robot)
8	*Slat*	*Size=3-hole*	*Assembled(11)*
9	Slat	Size=3-hole	Table(User)
10	Slat	Size=5-hole	Table(User)
11	*Comp(front)*	*Parts=(1,5,8)*	*UserHand*

Completing a step also affects the information state. In this case, the user has just built a component called a 'front', so we can update the model of the user's knowledge to indicate that this is likely to be a 'known' object. If a subsequent assembly task also requires a 'front', we can ask the user to build it without needing to explain it in detail, or we can ask the user if they remember the procedure.

5 Discussion

We have described the issues involved in representing assembly plans for use in a task-based dialogue system and shown how we use AND/OR graphs to represent Baufix assembly plans within the JAST human-robot dialogue system. We have then shown how the dialogue manager uses information from the task planner and the object inventory to describe the task plan and the required steps, to respond to actions and requests of the user, and to update the system state following assembly operations.

The dialogue manager has two distinct strategies available for describing a plan. With the top-down strategy, the structure of the plan is described before it is executed; with the depth-first strategy, the dialogue manager proceeds directly to concrete assembly operations and names sub-components only after they are completed. The current JAST system will shortly undergo a user evaluation in which naïve users interact with the system in the current robot-as-instructor scenario. Among other questions, this evaluation will compare these two strategies using measures such as user satisfaction and enjoyment and the success and efficiency of the assembly task.

The system is still under development, and several enhancements are planned for the next version. First, we aim to extend the system to support interactions in which both the robot and the user know the assembly plan. In such scenarios, it is likely that there would be much less verbal interaction between the participants. To support this, we will integrate components from another system (Erlhagen et al., 2007) that addresses a similar human-robot joint assembly task, but that uses dynamic neural fields to infer the user's goals from their non-verbal behaviour and to select complementary actions.

We would like to to move beyond the current small set of simple assembly plans, which are at the moment stored as hard-coded 'recipes' and loaded on request. It would increase the system's flexibility if an AND/OR graph could be created automatically or semi-automatically from a symbolic description of the assembled object; this would also enable the system to learn assembly plans interactively in cooperation with the user. More complex plans could also require different interaction strategies and a different, more flexible connection between the task planner and the dialogue manager in which a single assembly sequence is not selected at the start.

6 Acknowledgements

This work was supported by the EU FP6 IST Cognitive Systems Integrated Project 'JAST' (FP6-003747-IP). We thank the Planning/Language Interest Group at the University of Edinburgh and the Londial reviewers for useful feedback.

References

N. Blaylock and J. Allen. 2005. A collaborative problem-solving model of dialogue. In *Proceedings, 6th SIGdial Workshop on Discourse and Dialogue*, pages 200–211.

C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo. 2004. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1(2):315–348. doi:10.1142/S0219843604000150.

O. Brock. 1993. *InterPlan—ein interaktives Planungssystem*. Diplomarbeit (Master's thesis), Technical University of Berlin.

C. Callaway, M. Dzikovska, C. Matheson, J. Moore, and C. Zinn. 2006. Using dialogue to learn math in the LeActiveMath project. In *Proceedings, ECAI 2006 Workshop on Language-Enabled Educational Technology*.

W. Erlhagen, A. Mukovskiy, F. Chersi, and E. Bicho. 2007. On the development of intention understanding for joint action tasks. In *Proceedings, 6th IEEE International Conference on Development and Learning*. doi:10.1109/DEVLRN.2007.4354022.

T. W. Fong, I. Nourbakhsh, R. Ambrose, R. Simmons, A. Schultz, and J. Scholtz. 2005. The peer-to-peer human-robot interaction project. In *AIAA Space 2005*.

B. J. Grosz, S. Weinstein, and A. K. Joshi. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2). ACL Anthology J95-2003.

B. Jung. 2003. Task-level assembly modeling in virtual environments. In *Proceedings of Computational Science and Its Applications (ICCSA 2003)*.

A. Knoll. 2003. A basic system for multimodal robot instruction. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Perspectives on Dialogue in the New Millennium*, volume 114 of *Pragmatics & Beyond New Series*, pages 215–228. John Benjamins.

S. Kopp, B. Jung, N. Lessmann, and I. Wachsmuth. 2003. Max – a multimodal assistant in virtual reality construction. *Künstliche Intelligenz*, 4(03):11–17.

W. Mann and S. Thompson. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text*, 3:243–281.

L. S. Homem de Mello and A. C. Sanderson. 1990. AND/OR graph representation of assembly plans. *IEEE Transactions on Robotics and Automation*, 6(2):188–199. doi:10.1109/70.54734.

M. Rickert, M. E. Foster, M. Giuliani, T. By, G. Panin, and A. Knoll. 2007. Integrating language, vision and action for human robot dialog systems. In *Proceedings of HCI International 2007*. doi:10.1007/978-3-540-73281-5_108.

G. Rickheit and I. Wachsmuth. 2006. *Situated Communication*. Mouton de Gruyter, Berlin.

G. Sagerer, C. Bauckhage, E. Braun, J. Fritsch, F. Kummert, F. Lömker, and S. Wachsmuth. 2002. Structure and process: Learning of visual models and construction plans for complex objects. In G. D. Hager, H. I. Christensen, H. Bunke, and R. Klein, editors, *Sensor Based Intelligent Robots*, pages 317–344. Springer.

D. Traum and S. Larsson. 2003. The information state approach to dialogue management. In J. C. J. Van Kuppevelt and R. W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer Academic Publishers.

A Continual Multiagent Planning Approach to Situated Dialogue

Michael Brenner
Albert-Ludwigs-Universität
Freiburg, Germany
brenner@informatik.uni-freiburg.de

Ivana Kruijff-Korbayová
DFKI GmbH
Saarbrücken, Germany
ivana.kruijff@dfki.de

Abstract

Situated dialogue is usually tightly integrated with behavior planning, physical action and perception. This paper presents an algorithmic framework, Continual Collaborative Planning (CCP), for modeling this kind of integrated behavior and shows how CCP agents naturally blend physical and communicative actions. For experiments with conversational CCP agents we have developed MAPSIM, a software environment that can generate multiagent simulations from formal multiagent planning problems automatically. MAPSIM permits comparison of CCP-based dialogue strategies on a wide range of domains and problems without domain-specific programming. Despite their linguistic capabilities being limited MAPSIM agents can already engage in fairly realistic situated dialogues. Our ongoing work is taking this approach from simulation to real human-robot interaction.

(1) Anne: ”Please give me the coffee, R2D2.”
(2) R2D2: ”Okay.”
(3) R2D2: ”Where is the coffee, Anne?”
(4) Anne: ”The coffee is in the kitchen.”
(5) R2D2: ”Thanks, Anne.”
(6) R2D2: ”Please open the kitchen door, Anne.”
(7) Anne: ”Okay.”
(8) Anne opens the kitchen door.
(9) R2D2: ”Thanks for opening the kitchen door, Anne.”
(10) R2D2 moves to the kitchen.
(11) R2D2 takes the coffee.
(12) R2D2 moves to the living room.
(13) R2D2 brings Anne the coffee.
(14) Anne: ”Thanks for bringing me the coffee, R2D2.”

Figure 1: Mixed-initiative dialogue between two artificial agents in MAPSIM (*Household* domain).

1 Introduction

When several agents are situated in a common environment they usually interact physically as well as verbally. Verbal interactions in such environments, i. e. situated dialogues, both *reflect* the past and *influence* the future physical behavior of the agents. As a result, situated dialogue is continually *interleaved* with behavior planning, physical action and sensing. But when and why do agents switch between these rather distinct activities? In particular, how is dialogue triggered by physical events and how, in turn, does it constrain them?

In this paper, we approach these questions from the perspective of *multiagent planning*. Specifically, we describe situated dialogue as Distributed Continual Planning (DCP) (DesJardins et al., 1999), i. e. as a process that integrates planning, acting and perception with communication. We present a novel DCP algorithm called Continual Collaborative Planning (CCP) and show how it can be used for situated dialogue modeling. Interestingly, the role of communication in CCP is twofold: A dialogue move can be part of the collaborative *planning* process; however, it is also the *execution* of a communicative action and, just like the execution of a physical action, it changes the “world” in ways that may lead to previously unforeseen changes in plans and, consequently, additional interactions. Since goals and plans of agents are continually revised, CCP models very dynamic interactions that naturally include mixed-initiative subdialogues and interleaved physical and communicative actions.

Approaches to situated dialogue can only be evaluated in environments where agents are actually situated, i. e. where they can not only communicate, but also perceive and act. Because we want to evaluate CCP (and related approaches) over a

wide range of application domains we have developed MAPSIM, a simulation environment that turns formal multiagent planning problems into multiagent simulations. Crucially, MAPSIM creates the simulation as well as a domain-specific lexicon for natural-language dialogue *automatically* when analyzing the planning domain. Since no domain-specific *programming* is needed MAPSIM can be used to quickly evaluate dialogue strategies on a wide range of domains and problems.

The paper is structured as follows: We first introduce our multiagent planning formalism and discuss its suitability for dialogue planning. Then we present the CCP algorithm. In the subsequent sections we describe MAPSIM and analyze CCP dialogues in several domains. In the final sections we discuss related work and indicate our ongoing efforts.

2 Multiagent Planning Formalism

Planning in dynamic multiagent environments means reasoning about the environment, about (mutual) beliefs, perceptual capabilities and the possible physical and communicative actions of oneself and of others. All of these elements can be modeled in the multiagent planning language MAPL (Brenner, 2008). In this section we introduce MAPL informally and discuss its suitability for dialogue planning; formal definitions can be found in (Brenner, 2008).

MAPL is a multiagent variant of PDDL (Planning Domain Definition Language), the de facto standard language for classical planning (Fox and Long, 2003). One important extension in MAPL is the use of multi-valued state variables (MVSVs) instead of propositions. For example, a state variable *color(ball)* would have exactly one of its possible *domain* values *red, yellow,* or *blue* compared to the three semantically unrelated propositions *(color ball red), (color ball yellow), (color ball blue)*, all or none of which could be true in a given STRIPS state. MVSVs have successfully been used in classical planning in recent years (Helmert, 2006), but they also provide distinctive benefits when used for dialogue planning.

Firstly, we can use MVSVs to model *knowledge* and *ignorance* of agents: if no value is known for a state variable it is *unknown* (contrast this with the closed world assumption of classical planning: what is not known to be true is *false*). This concept can also be extended to beliefs about other agents' beliefs and mutual beliefs which are modeled by so-called **belief state variables**. Secondly, *wh-questions* can be modeled as queries about MVSVs in our model (see below). Thirdly, algorithms for generating referring expressions, such as the full brevity algorithm of (Dale, 1992), can be directly implemented using a MVSV representation.

(1) Bill goes home.
(2) Bill: "Please bake the pizza, Oven."
(3) Oven: "Okay."
(4) Oven bakes the pizza.
(5) Oven: "I have finished baking the pizza, Bill."
(6) Bill: "Thanks for baking the pizza, Oven."
(7) Bill: "Please bring me the pizza, R2D2."
(8) R2D2: "Okay."
(9) R2D2 brings Bill the pizza.
(10) Bill: "Thanks for bringing me the pizza, R2D2."
(11) Bill eats the pizza.

Figure 2: Dialogue between three artificial agents in MAPSIM (*Pizza* domain).

MAPL **actions** are similar to those of PDDL. In MAPL, every action has a **controlling agent** who executes the action and controls when it is done. Agents are fully autonomous when executing actions, i. e. there is no external synchronization or scheduling component. As a consequence an action will only be executed if, in addition to its preconditions being satisfied, the controlling agent *knows* that they hold. Implicitly, all MAPL actions are extended with such **knowledge preconditions** (cf. also (Lochbaum, 1998)). Similarly, there are implicit **commitment preconditions**, intuitively describing the fact that an agent will only execute actions if he has agreed to do so.

A MAPL domain can define three different ways to affect the beliefs of agents (necessary, e. g., in order to satisfy knowledge preconditions): sensing, copresence (joint sensing), and communication. All three are MAPL actions that have knowledge effects. **Sensor models** describe the circumstances in which the current value of a state variable can be perceived. **Copresence models** are multiagent sensor models that induce mutual belief about the perceived state variable (Clark and Marshall, 1981).

Informally, agents are copresent when they are in a common situation where they can not only perceive the same things but also each other. Individual and joint sensing are important for dialogue because they help *avoiding* it: an agent does not need to ask for what he senses himself, and he does not need to verbalize what he assumes to be perceived by the other agents as well. Communicative acts currently come in two forms: (i) **Declarative statements** are actions that, similarly to sensory actions, can change the belief state of another agent in specific circumstances. Line 5 of Fig. 2 shows an example of an agent explicitly providing another one with factual information. (ii) **Questions**, **commands** and **acknowledgments** are not explicitly modeled in a MAPL domain, but generated during CCP (as discussed in Sect. 3). These communicative acts potentially cover a broad range of speech acts, whose differentiation requires further refinement of the corresponding preconditions and effects.

MAPL **goals** correspond to PDDL goal formulae. However, MAPL has two additional goal-like constructs: **Temporary subgoals** (TSGs) are mandatory, but not necessarily permanent goals, i. e. they must be satisfied by the plan at some point, but may be violated in the final state. **Assertions**, on the other hand, describe *optional* "landmarks", i. e. TSGs that may helpful in achieving specific effects in later phases of the continual planning processes, which cannot be fully planned for yet because of missing information (Brenner and Nebel, 2006; Brenner, 2008). For example, the MAPL domain used to create the simulation in Fig. 1 contains an assertion stating that, informally speaking, to get something one must first know where it is.

MAPL plans differ from PDDL plans in being only *partially ordered*. This is inevitable since we assume that there is no central executive which could guarantee a totally ordered execution. We use the term **asynchronous plans** since MAPL plans also allow for *concurrent* occurrence of actions. Fig. 3 shows an example. An asynchronous plan that guarantees that the implied knowledge preconditions will be satisfied during execution (e. g. by explicitly naming the perceptions to be made and speech acts to be used) is called **self-synchronizing plan** because it "explains" how the agents can coordinate their behavior during execution.

It is often impossible for a group of situated agents to jointly commit to a self-synchronizing plan prior to beginning its execution. As an example, line 1 of Fig. 2 shows how an agent must start executing its individual multiagent plan (i. e. a plan for a group of agents but to which no other agent has committed yet) in order to even get the chance to negotiate the plan with the others: In this scenario, Bill must physically move first because he can only communicate with his household appliances "at home". This is modeled explicitly in the MAPL domain by means of a so-called **communication precondition** that the planner has to satisfy if agents should engage in dialogue. In future work, we will also use communication preconditions to model dialogue-specific requirements like *attention* (Grosz and Sidner, 1986) and *engagement* (Sidner et al., 2005).

3 Continual Collaborative Planning

Continual Collaborative Planning (CCP) agents switch between planning, partial plan execution, monitoring, plan adaptation and communication. Alg. 1 gives a high-level description of the CCP algorithm. Since the current state of the algorithm not only depends on what the agent has been doing, but also on the messages received from others, CCP is specified as a Distributed Algorithm (Lynch, 1996).

Algorithm 1 CCP AGENT(S, G)

$P = \emptyset$

Received no message:
  if S satisfies G do
    return "goal reached"
  else
    $P =$ MONITORINGANDREPLANNING(S, G, P)
  if $P = \emptyset$ then
    return "cannot achieve goal G"
  else
    $(S, P) =$ EXECUTIONANDSTATEESTIMATION(S, P)

Received (tell-val vx) **from** agent a:
  add $v \doteq x$ to S

Received request(e) **from** agent a:
  $sg =$ TRANSLATEREQUESTTOGOAL(e)
  $P =$ MONITORINGANDREPLANNING$(S, G \cup sg, \emptyset)$
  if $P = \emptyset$ then
    send "cannot execute request e" **to** a
  else
    add sg to G as temporary subgoal

We will first discuss the base case when no communication has taken place yet, i. e. the CCP agent has neither sent nor received any messages yet.

Roughly speaking, the agent alternates between (re-)planning and acting in this case. The two phases are detailed in Algs. 2 and 3. Alg. 2 shows how a new planning phase is triggered: the agent *monitors* whether his current plan has become invalid due to unexpected (external) events or changes in his goals. If this is the case, the agent adapts its plan by replanning those parts that are no longer executable. In order to exploit the power of state-of-the-art planning systems, Alg. 2 uses an unspecified classical planner PLANNER to (re-)plan for the obsolete or missing parts of the old plan. The details of this process are irrelevant for the purpose of this paper; it results in an asynchronous plan that specifies actions for (possibly) several agents and the causal and temporal relation between them that is necessary to achieve the planning agent's goal.

Algorithm 2 MONITORINGANDREPLANNING(S, G, P)

```
if res(S,P) ⊉ G
   REMOVEOBSOLETESUFFIXGRAPH(P)
   P' = PLANNER(A, res(S,P), G)
   P = CONCAT(P, P')
return P
```

Fig. 3 shows such an asynchronous plan for the *pizza* scenario of Fig. 2, created with Alg. 2. Note that this plan contains special *negotiation* actions; they will be the triggers for task-orientated subdialogues in a later phase of CCP. The planning algorithm enforces such negotiation actions to be included in a plan whenever this plan includes actions or subplans to be executed not by the planning agent, but by another agent who is not yet committed to this plan. Thus CCP ensures that a (sub-)dialogue will take place that either secures the other agent's commitment or triggers replanning. Note how, in turn, the need for negotiation has forced the planner to include a physical action (Bill's moving home) into the plan in order to satisfy the abovementioned communication precondition.

As soon as a CCP agent has found (or repaired) a valid plan it enters the execution phase, described in Alg. 3. First, an action, e, on the first level of the plan, i. e. one whose preconditions are satisfied in the current state, is chosen non-deterministically. If the action is controlled by the CCP agent himself, it is executed. If not, the planning agent tries to determine whether the action was executed by its controlling agent. In both cases, the CCP agent will try to update its knowledge about the world state based on the expected effects and the actual perceptions made (FUSE function).

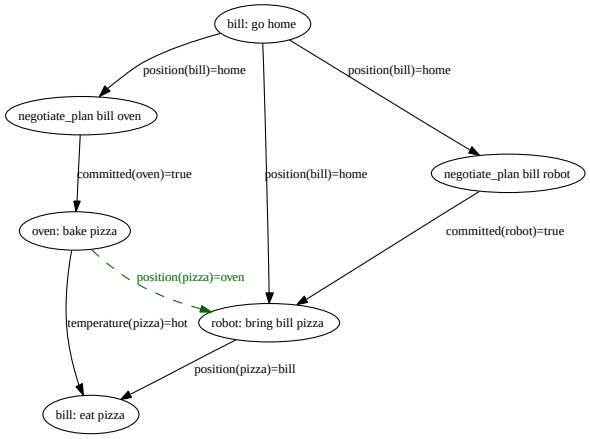


Figure 3: Bill's plan for getting pizza.

Algorithm 3 EXECUTIONANDSTATEESTIMATION(S, P)

```
e = choose a first-level event from P
if e ='negotiate_plan with agent a'
   r = SELECTBESTREQUEST(P, a)
   send request(r) to a
else if agt(e) = self then
   EXECUTE(e)
S' = app(S, e)
exp = EXPECTEDPERCEPTIONS(S', A^s)
perc = GETSENSORDATA()
if perc ⊇ exp or exp = ∅ then
   remove e from P
S = FUSE(S', perc)
return (S,P)
```

The most interesting case for this paper is the one where the action chosen to be executed is *negotiate_plan*. This means that a CCP agent (A) is now in a situation where he is able communicate with another agent (B) that he intends to collaborate with, i. e. A's plan includes at least one action controlled by B, that B has not yet committed to. In this case, A will send a *request* to B. However, if a plan contains several actions by another agent, i. e. a whole subplan, it is often best not to request execution of the actions individually, but to ask for the end result or, respectively, the final action in the subplan. In other situations it may even be reasonable to request the achievement of subplans that include more than one agent. CCP does not stipulate a specific implementation of SELECTBESTREQUEST; we will describe

one version in Sect. 5.

When an agent receives a request, Alg. 1 enters into a new phase. First the request is translated into a goal formula (Brenner, 2007) and tested for achievability. This is a simplification for the sake of processing efficiency, based on the assumption that what matters to the other agent is not the exact action, but its *result*, i. e. the achievement of a goal or precondition for a subsequent action by the requesting agent. Additionally, constraints on the arguments of requests (e. g. intended referents of natural language expressions) are easier to model as goal constraints than as actions (Brenner, 2007). Accepted requests are adopted as *temporary subgoals* (TSGs). This means that they must only be achieved temporarily and do not have to hold any more when the agent's main goal is achieved.

The adoption of requests as TSGs is a crucial element of CCP that, to the best of our knowledge, has not been described in other Continual Planning approaches: in addition to repeatedly revising their beliefs about the world, CCP agents also perform continual *goal revision*. In the simplest case, this leads to information-seeking *subdialogues*, as in lines 3–5 of Fig. 1. But newly adopted TSGs also explain why agents engage in subdialogues that mix communicative and physical actions (as in lines 6–9 of the same example).

4 MAPSIM

Continual Planning approaches can only be tested in environments where agents can actually execute, monitor and revise their plans. This is all the more true for our DCP approach to situated dialogue where agents need to interact collaboratively. To this end we have developed MAPSIM, a software environment that automatically generates multiagent simulations from MAPL domains. In other words, MAPSIM interprets the planning domain as an *executable model* of the environment. Thus, MAPSIM allows designers of DCP algorithms to evaluate their approaches on various domains with minimal effort. In this section, we give an overview of MAPSIM and describe how it is used for generating situated dialogues.

The MAPL domain description is parsed, analyzed and turned into perception, action, and communication models for CCP agents. During the simulation, MAPSIM maintains and updates the global world state and it uses the sensor models to compute individual and joint perceptions of agents. The agents interact with the simulation by sending *commands* in the form of plain MAPL actions. The simulator then executes the action, i. e. it checks the preconditions and applies effects as specified in the MAPL domain. If the controlling agent of a command is not identical to the agent who sent it to the simulator this is interpreted as a *request* which is not directly executed but passed on to the corresponding agent. MAPSIM also accepts specific commands for acknowledging subgoal acceptance and subgoal achievement.

```
(1)  Anne: request R2D2 'give R2D2 coffee Anne'.
(2)  R2D2: accept_request 'give R2D2 coffee Anne'.
(3)  R2D2: request Anne 'tell_val Anne R2D2 pos(coffee)'.
(4)  Anne: execute 'tell_val Anne R2D2 pos(coffee)'.
(5)  R2D2: ack_achieved 'tell_val Anne R2D2 pos(coffee)'.
(6)  ...
```

Figure 4: The MAPSIM run of Fig. 1 without NL verbalization.

Agents do not need to know anything about how their actions are executed. Thus, they can implement arbitrary deliberative or reactive methods to determine their behaviour and their reactions to requests. We believe that this can make MAPSIM a valuable evaluation tool even when the DCP and dialogue strategies investigated differ significantly from CCP. For example, the simulated dialogues produced by MAPSIM using different strategies could be evaluated using objective measures such as task success or dialogue costs from the PARADISE framework (Walker et al., 1997).

MAPSIM and the CCP agents described in this paper have been implemented in Python, using state-of-the-art planning technology as subsolvers.The generic planner currently used for CCP is a slightly modified version of Axioms-FF (Thiebaux et al., 2003). This enables MAPSIM to generate dialogues between artificial agents very fast.[1]

The main goal of this work is to show how a generic multiagent planning algorithm can be used

[1]For example, during the dialogue of Fig. 1 CCP called the PLANNER function 13 times with a total planning time less than half a second on a 1.6 GHz AMD Athlon.

for situated dialogue in natural language, e. g. in human-robot interaction (HRI). It is therefore important to investigate the efforts needed for mapping between the MAPL-based representation used by CCP agents and natural language. To that end MAPSIM includes a verbalization module, called the *reporter* agent. The reporter observes all physical and communicative events in the simulation and verbalizes them in English. All dialogues shown throughout the paper are unaltered outputs of the reporter. Fig. 4 shows the beginning of the MAPSIM run of Fig. 1 with reporting turned off.

The reporter is a simple template engine that first determines an appropriate pattern depending on the command type currently executed, then recursively replaces templates with concrete arguments until a template-free sentence is generated. Base values for arguments are generated directly from analyzing the MAPL domain. For example, operator names are assumed to directly correspond to verbs. Standard templates can be overridden by domain-specific patterns. However, the only general need for this we have experienced is the definition of verb complements. For example, the *Household* domain defines the complement of *"move"* as *"to the $arg0"* where *$arg0* is instantiated with the first argument of the respective command. Apart from verb complements, the *only* domain-specific template that was necessary to generate Fig. 1 states that the *interrogative* (wh-word) for state variables *position(x)* is *"where"*. While, compared to "real" natural-language processing systems, this is a simplistic approach with obvious limits, the minimal effort needed to achieve fairly realistic surface generation is noteworthy and will be exploited in future work.

5 Detailed Analysis of MAPSIM runs

This section provides a detailed analysis of several CCP runs in MAPSIM. It is important to realize that none of the sample runs in this paper is based on just one multiagent plan, but on a *series* of plans, devised, partly executed and revised several times according to Alg. 1.

All dialogue in CCP is driven by individual desires, i. e. agents engage in dialogue only if they need help in satisfying their individual goals. In the *household* scenario (Fig. 1) the necessity for collaboration stems from the fact that only R2D2 can move to the kitchen to get coffee, but only Anne can open the kitchen door. In the *pizza* scenario (Fig. 2) Bill needs the collaboration of his intelligent household appliances to be able to eat pizza.

As we have already seen, Bill's initial individual planning process resulted in the multiagent plan shown in Fig. 3. In this situation, Alg. 3 can only choose a *physical* action for Bill to execute, namely *go home*. Note that only the *execution* of this action enables Bill to subsequently communicate with Oven and R2D2 at all. Thus, Bill's problem can *only* be solved by a DCP approach that is able to interleave planning, physical execution and dialogue whenever necessary.

When at home, Bill can (and must) negotiate his plan with the two other agents he wants to involve. Alg. 3 uses the black-box function SELECTBESTREQUEST to determine an appropriate *temporary subgoal* whose achievement will be requested from another agent.

The currently used REQUESTSUBPLAN strategy works as follows: the agent first determines the longest possible subplan involving only one agent, then chooses an action on the *final* level of this plan as the best request. In other words, a CCP agent posing a request does not specify details about how he wants a temporary subgoal to be achieved. In the *household* example, Anne thus does not request R2D2 to go to the kitchen and get the coffee there, but just requests the last action in her multiagent plan, namely the robot giving her the coffee.

Admittedly, the straightforward verbalization of this action by the reporting agent using the verb "give" results in an unnatural dialogue contribution. Anne's request would be more appropriately formulated using "bring", "fetch" or "get", which unlike "give" do not presuppose that R2D2 already has the coffee. This reveals the need to take more of the subplan into account when verbalizing the request, a topic we are taking up in further work.

Anne thus leaves it to R2D2 to find its own solution to achieve the TSG. This "lazy" strategy mirrors on the dialogue level the idea of the Continual Planning approach, where an individual CCP agent postpones the solution of some subproblems to later phases in the planning-execution-monitoring cycle.

R2D2's previous plan was to do nothing (which satisfied his "empty" goal). After adopting the new

TSG, this plan is no longer valid and Alg. 2 triggers a new planning phase. Since R2D2 does not know where the coffee is this plan includes an appropriate information-gathering action and postpones detailed planning for getting the coffee until this information is known (by means of an *assertion* (Brenner and Nebel, 2006)). In our example, the information-gathering action is a request for information to Anne (cf. line 3 of Fig. 1). This request is generated as follows: R2D2's plan contains the action *(tell-val Anne R2D2 pos coffee)*, i. e. a speech act to be executed by Anne. According to Alg. 3, this action to be performed by another agent (from R2D2's point of view) must be requested first. Line 3 of Fig. 4 shows this request when executed without the *reporter* agent. Its verbalization results in R2D2 asking the question "Where is the coffee, Anne?".

MAPSIM provides several options for generating *acknowledgments*. In the dialogues presented here, agents provide acknowledgments when they accept a request (e. g. lines 2 and 7 of Fig. 1) and also when they realize that a request of theirs has been satisfied (e. g. lines 5, 9 and 14 of Fig. 1). Note that answers to questions are acknowledged only briefly, but satisfaction of physical subgoals is acknowledged more explicitly. While this is not necessarily the best acknowledgment strategy, it shows how the multiagent plan and the CCP history provide *context* as well as *focus* (Grosz and Sidner, 1986)) that can easily be exploited for surface generation and, in the future, also for interpretation (cf. Sect. 7).

For lack of space, we cannot discuss the rest of the dialogues in detail. Note, however, how the agents switch seamlessly between communicative and physical actions whenever necessary. Not shown by the reports are the *perceptions* made by the agents during the runs. Nevertheless, they are important for the dialogue, too, since agents also reason about their mutual perceptions and thus can avoid unnecessary verbalizations.

6 Related Work

This work shares many characteristics with previous approaches modeling dialogue as *collaborative planning*, most notably those based on the SharedPlans formalism (Grosz and Sidner, 1990; Grosz and Kraus, 1996; Lochbaum, 1998). SharedPlans use much more elaborate mental attitudes than MAPL and CCP, mainly because CCP agents rely on them only *implicitly* – until a violation of their assumptions prompts plan adaptation or new dialogue. In this respect, the commitments made by CCP agents more resemble the *joint persistent goals* of (Cohen and Levesque, 1991). Nevertheless, SharedPlans can be regarded as a "specification" of the kind of collaboration CCP intends to model *computationally*.

(Blaylock et al., 2003) note that SharedPlans do not model the cooperation that occurs during *execution*. They propose a high-level model of dialogue as *collaborative problem solving* (CPS); our approach can be regarded as an instantiation of that model. However, our work complements both SharedPlans and CPS by describing *how* knowledge preconditions prompt active sensing and information gathering during situated dialogue.

Distributed Continual Planning has been advocated as a new paradigm for planning in dynamic multiagent environment (DesJardins et al., 1999). To the best of our knowledge, ours is the first principled attempt to apply DCP to dialogue planning and also the first DCP approach describing deliberative *goal revision* as part of a DCP algorithm.

Collagen (Rich et al., 2001) is a system for building collaborative interface agents that is based on (Grosz and Sidner, 1986; Grosz and Sidner, 1990), which is domain-independent and has been used for various applications. Collagen's methods for representing the discourse state and doing plan recognition are much more sophisticated than CCP currently. However, Collagen does not (yet) include a first-principles planner, but relies on plan libraries and domain-specific code plug-ins (Rich and Sidner, 2007). It would be interesting to investigate whether CCP can be integrated with Collagen.

Similarly, the most prominent representative of the information-state-update approach to dialogue modeling, GoDiS (Traum and Larsson, 2003), has complementary rather than competing main strengths: GoDiS has a more elaborate repertoire of dialogue moves and can produce more sophisticated dialogue behavior than CCP and MAPSIM, but it uses static plans, and it is not clear how it would combine communication with physical action.

7 Conclusion and Outlook

We have presented a new algorithmic framework in which situated dialogue is modeled as Continual Collaborative Planning (CCP). We have shown how mixed-initiative dialogue that interleaves physical actions, sensing, and communication between agents occurs naturally during CCP. As a practical contribution, we have developed MAPSIM, a software tool that automatically generates multiagent simulations from formal planning domains, thus permitting the evaluation of CCP and other dialogue strategies on a wide range of applications.

The questions raised in the introduction about when and why agents switch between planning, acting, and execution have, intuitively, been answered as follows by CCP: Agents (re)start planning as soon as their plan becomes obsolete, possibly not because the world, but because their *goals* changed. They act whenever they have a valid plan containing executable physical actions. And they engage in dialogue whenever they want others to share subgoals or are requested to do this themselves. Since situated communication may have (physical) preconditions that must be satisfied first (e. g. being in the same room, having the other agent's attention/engagement, etc.) CCP explains how the need for dialogue may also trigger additional planning and acting.

From simulation to human-robot interaction The work presented in this paper provides a starting point for developing agents, e. g. robots, that can engage in situated dialogue with humans. Indeed, we are currently implementing CCP on a robotic system in the CoSy project. To that end, we are extending our approach in the following respects: (1) To allow for imperfect communication, we need to improve the handling of acknowledgments to include positive as well as negative feedback and clarifications. (2) To support the full range of plan-negotiation between dialogue participants, we need to allow agents to reject requests and accept rejections from others. This will enable us to handle situations with, e. g., conflicting goals, discrepancies in beliefs and execution failures.

Doing this amounts to refining and extending the repertoire of speech acts. Since the planning technology underlying CCP is known to scale very well (Thiebaux et al., 2003), we expect our dialogue approach to also scale up well to a larger repertoire of speech acts, more complex interactions and higher numbers of interacting parties.

We are also investigating how to better expose the purpose that an individual dialogue move serves in achieving an agent's overall goals, e. g. by deriving an explicit *dialogue plan* during CCP. Such a plan, in combination with the current state of the CCP process, will provide rich context information to the linguistic components of our robot, e. g. for the task of utterance interpretation and contextually appropriate surface generation.

Acknowledgments

This work has been supported by the EU in the Integrated Project "CoSy" (FP6-004250).

References

N. Blaylock, J. Allen, and G. Ferguson. 2003. Managing communicative intentions with collaborative problem solving. In *Current and New Directions in Dialogue*. Kluwer.

M. Brenner and B. Nebel. 2006. Continual planning and acting in dynamic multiagent environments. In *Proc. PCAR*, Perth, Australia.

M. Brenner. 2007. Situation-aware interpretation, planning and execution of user commands by autonomous robots. In *Proc. IEEE RO-MAN 2007*.

M. Brenner. 2008. The multiagent planning language MAPL. Technical report, Albert-Ludwigs-Universität, Institut für Informatik, Freiburg, Germany.

H. H. Clark and C. R. Marshall. 1981. Definite reference and mutual knowledge. In *Elements of discourse understanding*. Cambridge University Press.

P. Cohen and H. Levesque. 1991. Teamwork. *Noûs*, 25(4).

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

M. DesJardins, E. Durfee, Jr. C. Ortiz, and M. Wolverton. 1999. A survey of research in distributed, continual planning. *The AI Magazine*.

M. Fox and D. Long. 2003. PDDL 2.1: an extension to PDDL for expressing temporal planning domains. *JAIR*.

B. J. Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86.

B. J. Grosz and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3).

B. Grosz and C. Sidner. 1990. Plans for discourse. In *Intentions in Communication*. MIT Press.

M. Helmert. 2006. The Fast Downward planning system. *JAIR*, 26:191–246.

K. E. Lochbaum. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*.

N. Lynch. 1996. *Distributed Algorithms*. Morgan Kaufmann, San Francisco, CA.

C. Rich and C. L. Sidner. 2007. Generating, recognizing and communicating intentions in human-computer collaboration. In *AAAI Spring Symposium on Intentions in Intelligent Systems, Stanford, CA*.

C. Rich, C. L. Sidner, and Neal Lesh. 2001. Collagen: applying collaborative discourse theory to human-computer interaction. *The AI Magazine*, 22(4).

C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. 2005. Explorations in engagement for humans and robots. *AIJ*.

S. Thiebaux, J. Hoffmann, and B. Nebel. 2003. In defense of axioms in PDDL. In *Proc. IJCAI*.

D. Traum and S. Larsson. 2003. The information state approach to dialogue management. In *Current and New Directions in Discourse and Dialogue*. Kluwer.

M. Walker, D. Litman, C. Kamm, and A. Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proc. ACL-97*.

Accommodation through Tacit Sensing

Luciana Benotti
TALARIS Team - LORIA (Université Henri Poincaré, INRIA)
BP 239, 54506 Vandoeuvre-lès-Nancy, France
Luciana.Benotti@loria.fr

Abstract

The aim of this paper is to use insights from the theory of accommodation to study in a uniform way different kinds of acts involved in situated dialogue. When interlocutors are engaged in situated dialogue, their informational states evolve through dialogue acts, physical acts and sensing acts. We model this evolution in a non-traditional conversational system using tools from mature branches of artificial intelligence. In particular, we use a planner that is able to find plans in the presence of incomplete knowledge and sensing: PKS (Petrick and Bacchus, 2002). In the resulting model, we study the interactions among dialogue acts, physical acts and sensing acts, and their relationship with accommodation.

1 Introduction

The phenomena of accommodation has been widely studied from philosophical and linguistic perspectives, ranging from classical papers like (Lewis, 1979) to recent contributions like (Beaver and Zeevat, 2007). We view accommodation theory as a schema in which to study, in a uniform way, the different kinds of acts that occur in situated dialogue. We not only believe that such an approach can help us obtain better models of dialogue, but also that dialogue is an essential setting in which to test such theories, theories that are too frequently divorced from the commonest setting of language use: situated conversation.

When interlocutors are engaged in situated dialogue, it is evident that their informational states evolve as a result of the dialogue acts performed during the task, and through the physical acts that interlocutors perform on their environment. But their states are also updated with the information that the participants sense from their environment; embedded agents do not have complete information about the world but they can sense it.

The approach presented here uses insights (and tools) from mature branches of artificial intelligence, such as planning with incomplete knowledge and sensing, in order to build a model for non-traditional conversational systems. In general, traditional conversational systems assume that conversational partners share common goals and collaborate in order to perform the task at hand as efficiently as possible. Our setup explores a case where conversational partners do not share a common goal and they are not as cooperative as partners involved in task-oriented dialogue. Our setup is a text-adventure game, where one of the participants is the player and the other participant is the game. The game has all the information needed to solve the game task but this is not its goal; its goal is to make the interaction engaging and challenging, encouraging the player to explore and discover the game world.

This work is part of a larger project on reconciling linguistic reasoning and collaborative reasoning in conversation (Benotti, 2007; Benotti, 2008). We advance this program here by adding to our model the treatment of sensing actions. To this end, we have integrated in a conversational system a planner that is able to find plans in the presence of incomplete knowledge and sensing: PKS (Petrick and Bacchus, 2002). In the resulting model, we study the interactions among dialogue acts, physical acts and sensing acts. We believe that this issue relates in relevant ways with the theoretical question: “In which contexts can sentences that have particular implicatures felicitously oc-

cur?" Following (Beaver, 1994) we believe this to be a better formulation of the problem of accommodation than the traditional question: "What inferences do people draw from sentences?"

2 Accommodating when talking, acting and sensing: everyday examples

Accommodation and grounding of dialogue and physical acts are topics that have been widely studied. But the study of accommodation and grounding of sensing acts is also essential when agents are embedded. Moreover, even when interlocutors are co-situated, sensing acts are usually less evident than physical and dialogue acts. Hence, an important question to study is "When is the common ground of the dialogue updated with the sensed information?" Or in other words, "When is there in the state of the activity enough evidence that a piece of information has been sensed?"

Let us address these questions with an example:

In kindergarden, the teacher showed a green square to a boy and, offering a piece of paper, told him: "Paint a circle that has this same color".

This simple example illustrates the interaction of a *dialogue act* performed by the teacher (request) with a *sensing action* (sense color) and a *physical action* (paint) that the teacher expects from the boy. When giving this instruction the teacher relied on the ability of the boy to sense the colors, but the sensing action is left tacit in the teacher request. She could have make it explicit saying "Look at the color of the square and paint a circle that has the same color". However in conversation, sensing actions are more naturally left tacit than made explicit. Why? Because they are so natural for sensing agents (indeed, sometimes they are unavoidable) that it is extremely easy to take them for granted.

Now we are going to look at this example as an instance of the general *rule of accommodation* introduced by Lewis in the article in which he coins the word *accommodation*:

If at time t something is said that requires component s_n of conversational score to have a value in the range r if what is said is to be true, or otherwise acceptable; and if s_n does not have a value in the range r just before t; and if such and such further conditions hold; then at t the score-component s_n takes some value in the range r. (Lewis, 1979, p.347)

This rule will help us perform a detailed analysis of our example in order to address the questions raised in the beginning of this section. Bearing this schema in mind, let us analyze step by step the different values that the variables of the rule take for our simple example. First of all, what's t? This is what Stalnaker has to say here:

The prior context that is relevant to the interpretation of a speech act is the context as it is changed by the fact that the speech act was made, but prior to the acceptance or rejection of the speech act. (Stalnaker, 1998, p.8)

So in our example t is the time right after the teacher said "Paint a circle that has this same color" but before the acceptance or rejection of this request.

Now, let us determine what the relevant components s_n are. Suppose that the boy is color blind and the teacher knows it. Then her request does not make much sense and any side participant and the boy himself will start asking what the goal of the request is, because clearly it cannot be the literal one: to obtain a green circle. Therefore, the color referred to by the teacher is the s_1 of our example. And if what the teacher said is to be acceptable, s_1 is required to have a particular value r_1; the same color than the square has in the real world (or in fact, a representation of it). Furthermore, there is no evidence that s_1 already has the value r_1 before the teacher began to speak (that is, there is no evidence that the color has been under discussion before), so we can assume that it doesn't.

Now, what are the further conditions that need to hold so that, at t, the score-component s_1 takes some value r_1? The teacher and the boy both know (at least intuitively) that people can sense their environment, that members of the same culture usually assign the same name to the same parts of the spectrum of colors, that humans can remember facts that they sense, that the sensed object is accessible, that a person will actually sense the color of an object if he is required to know this fact; the teacher and the boy rely on these and many other things that are usually taken for granted. All this knowledge is necessary for the boy to come up with the right sequence of actions in order to respond to the teacher's request; that is, in order to sense the color of the square and paint the circle.

Following Lewis, we would finish our instantiation of the rule of accommodation with the fact

that at the time t the score-component s_1 takes value r_1. Two last comments are in order here. First, it is worth pointing out that at the moment t the request has not yet been accepted or rejected but the addressee has already taken it in and adjusted himself to the fact that the dialogue act has been performed. The acceptance or rejection can be seen as a second change to the conversational record that ocurrs after the rule of accommodation applies. It's very important to distinguish between these two changes. Why? Because even if the request is rejected, the update of the conversational record that resulted from the accommodation may remain. Even if the boy answers "I don't like green. I won't do it", we know that the boy sensed the color of the square.

Second, how does the score-component s_1 takes value r_1? This is a question that is not directly addressed by Lewis but he seems to suggest is that s_1 takes value r_1 and nothing else changes. However, we agree with (Thomason et al., 2006; Hobbs et al., 1993; Kreutel and Matheson, 2003) that what is accommodated in order for s_1 to take value r_1 could be much more than just this fact. If we claimed that only s_1 changes, how can we explain the fact that the boy may take off a blindfold (he was playing "Blind man's bluff") after hearing the teacher? The required updates can also have their requirements (or preconditions) and side-effects, and we think that a natural way to model the accommodation updates is through *tacit acts*.

We adhere then to the view that explains Lewis' broad notion of accommodation (not limited to classical cases of presupposition accommodation) as tacit acts. Physical acts can be left tacit (Benotti, 2007), dialogue acts can be left tacit (Kreutel and Matheson, 2003; Thomason et al., 2006), but also sensing acts can be left tacit (this paper, Section 4). This is not a new idea then, but it's a promising approach and needs to be further developed.

The analysis of our example so far has given us some insight on the questions that were raised in the beginning of this section. We have seen that tacit sensing can be grounded even if the dialogue act that required the sensing is directly rejected (the "boy doesn't like green" example). And it can also be the case that the tacit sensing is grounded even if it cannot be directly executed because, for instance, it requires the execution of some physical act first (the "Blind man's bluff" example). The interactions among sensing acts, dialogue acts and physical acts can be extremely subtle; modelling them (putting sensing, physical and dialogue acts in a common schema) and, in particular making explicit the information at play, is the topic of the rest of this paper.

But first, let us have a look at a few more everyday examples; the aim of these instances is to show how frequent and pervasive are the interactions among different kinds of acts.

2.1 Tacit sensing and referring

If referring is treated as a dialogue act on its own, as many current dialogue systems do (DeVault and Stone, 2006), then the interaction between tacit physical action, tacit sensing action and referring acts need to be controlled. Consider this example:

Suppose that you are told that the hidden treasure you are seeking is behind the blue door. Painting a door blue does not satisfy the goal of finding the blue door — it merely obscures the entity of the appropriate door. (Etzioni et al., 1992, p.116)

This is an example of the incorrect interpretation (painting a door blue) that a conversational system can assign to a command when the system does not have complete information about the environment and has no restrictions on the order in which it can execute actions.

A first conclusion given this observation would be that only sensing actions (and not physical actions) should be allowed before referring actions are resolved. However, it might be the case, for example, that the blue door is in a different room, so the physical action of moving should be allowed before resolving the reference. A more refined approach would be then to leave the relevant properties of the definite description unchanged until the referred object is found. Current off-the-shelf planners provide ways in which to represent properties that must not change (usually called hands-off properties). Using this it is possible to model the fact that searching for a blue door is legitimate, whereas painting some door blue is not.

2.2 Tacit grounding and sensing

During dialogue, grounding acts are frequently left tacit. Consider the following example:

A[1]: Helen did not come to the party.
B[2]: How do you know that?
A[3]: Her car wasn't there.

B[4]: She could have come by bicycle.
(Kreutel and Matheson, 2003, p.6)

In this example, [4] tacitly grounds the assertion [3] (Helen's car wasn't there) but [4] also rejects the fact that [3] is a reason for [1]. In other words, [4] performs two dialogue acts that can be made explicit with "[4']: Ok, her car wasn't there. She could have come by bicycle". Notice that [4] cannot tacitly reject the assertion [3], something is wrong with: "??I saw her car there. She could have come by bicycle".

Sensing actions offer a whole new world for tacit grounding in situated interaction. After giving an instruction, for example, just sensing the results of the required acts is often the best way to know whether the addressee understood (and hence, grounded) the instruction. If you tell your daughter "Turn off the light of your room" and, when you come back, the light is off then you are pretty sure that she heard you.

2.3 Sensing and tacit exogenous events

Unobserved exogenous events can change the value of properties that have already been sensed. So we may well be faced with the treatment of not only incomplete but also incorrect information. But if we assume that the state of the world evolves via the effects of actions and events, then there is a intuitive approach for updating sensed values. Whenever a sensed property needs to be updated in order to make sense of the evolution of the interaction, a tacit exogenous event that updates this property can be inferred. In the following example, Andrew might have sensed that the contents of the pot were raw, but after a while he observe Bess's actions and update his knowledge.

Perhaps only Bess will see when the contents of her boiling pot have cooked. Andrew might still infer that this event has taken place from observing Bess's actions — say, by watching Bess turn off the heat or empty the pot. (Thomason et al., 2006, p.16)

In this three subsections we have shown everyday examples of the interaction of sensing acts, physical acts and dialogue acts. But these are only the tip of the iceberg. We believe that research on such interactions will be fundamental to deepening our understanding of situated dialogue. But how can we model these interactions? This is the topic of the next two sections.

3 A technical framework for tacit sensing

In this section we will introduce the two systems used to implement the ideas discussed in the previous section. We first briefly present our conversational application (the text adventure game), and then describe the main features of the planner that we use for our formalization and case-studies.

3.1 Situated interaction in a text-adventure

We have implemented a text-adventure game which can interpret commands that require tacit sensing. In this game-engine, the player can be embedded in different simulated game environments. The player can issue natural language requests to the game in order to manipulate and change the game environment. She can request to *sense the game objects* through special actions such as read; or directly *perceive the environment* (for example, every time the player enters a new room the game describes it). The situated perspective, and the answers generated by the game as a result of the player requests, allow the player to discover the rules by which her environment is governed and to extend her knowledge accordingly.

A game scenario is represented by several informational components: a database that specifies STRIPS-like actions (Fikes et al., 1972), a grammar, a lexicon, and two description logic knowledge bases (Baader et al., 2003) that share a set of definitions (one knowledge base models the player knowledge and the other the game scenario). The natural language processing module receives the player command and outputs a flat semantic representation that is used by the action handling module to modify the game scenario. The natural language generation module verbalizes the results of the player actions. (Benotti, 2007) describes how classical planning capabilities can be integrated into the architecture of the game-engine. Such planning abilities allow the game to infer *physical actions* left tacit by the player using the off-the-shelf planner Blackbox (Kautz and Selman, 1999). Blackbox implements classical planning techniques and assumes complete knowledge about the planning domain. In this paper, the planner PKS (Planning with Knowledge and Sensing) is used in order to investigate the case in which *sensing actions* are tacit.

3.2 Planning with knowledge and sensing

PKS (Petrick and Bacchus, 2002) is a knowledge-based planner that is able to construct conditional plans in the presence of incomplete knowledge. PKS builds plans by reasoning about the effects of actions on an agent's knowledge state, as opposed to other approaches based on possible-world reasoning. By reasoning at the knowledge level, PKS can avoid some of the irrelevant distinctions that occur at the world level, improving efficiency and producing natural plans. The PKS specification language offers features such as functions and variables, allowing it to solve problems that can be difficult for traditional planners (and making it ideal for non-traditional dialogue systems).

PKS is based on a generalization of STRIPS. In STRIPS, the world state is modelled by a single database. In PKS, the planners knowledge state, rather than the world state, is represented by a tuple $\langle K_f, K_w, K_v, K_x \rangle$ of databases whose contents have a fixed, formal interpretation in epistemic logic. Actions are specified as updates to these databases using the knowledge primitives *know fact* which modifies the database K_f, *know value* which modifies the database K_v, *know whether* which modifies the database K_w, and *know which* which modifies the database K_x.

We briefly describe these four databases here. K_f is like a standard STRIPS database except that both positive and negative facts are stored and the closed world assumption does not apply. K_v stores information about function values that will become known at execution time, such as the plan-time effects of sensing actions that return numeric values. K_w models the plan-time effects of binary sensing actions that sense the truth value of a proposition. K_x models the agent's exclusive disjunctive knowledge of literals (that is, the agent knows that exactly one literal from a set is true).

PKS performance has been tested for the composition of web services with promising results (Martinez and Lesperance, 2004). Moreover, in the prototype we have implemented using PKS inside our text-adventure game, PKS response time was acceptable (less than 2 seconds) for the kind of planning problems that the text adventure typically gives rise to. We tested it using the breadth first search strategy, rather than depth first because we require optimal length plans.

4 Tacit sensing: 2 case-studies

In this section we are going to explain in detail how a command issued by the player that includes tacit sensing actions is interpreted using PKS, and then executed by the game. We first classify sensing actions as either disjunctive or existential. We then present a case-study of disjunctive knowledge that makes use of conditional plans. Finally, we describe a case-study of existential knowledge that makes use of parametric plans.

4.1 Incomplete knowledge and sensing

There are two sorts of sensing actions, corresponding to the two ways an agent can gather information about the world at run-time. On the one hand, a sensing action can observe the truth value of a proposition $P(c)$, resulting in a conditional plan. The kind of incomplete knowledge sensed by this kind of action can be described as *binary* because it represents the fact that the agent knows which of the two disjuncts in $P(c) \vee \neg P(c)$ is true. In PKS, binary sensing actions are those that modify the K_w knowledge base. On the other hand, a sensing action can identify an object that has a particular property, resulting in a plan that contains run-time variables. The kind of incomplete knowledge sensed by these kind of action can be described as *existential* because it represents the fact that the agent knows a witness for $\exists x.P(x)$. In PKS, existential sensing actions are those that modify the K_v database.

We will now explain in detail how these two kinds of sensing actions can be left tacit by the player in our text-adventure game.

We said that our model can handle incomplete knowledge about the interaction in which a dialogue is situated. But how incomplete is the knowledge the model can handle? There are several levels at which knowledge can be incomplete. The most studied scenario is one in which not all the properties and relations of the objects involved in the task are known, but the set of objects is finite and all objects are named (that is all objects are associated with a constant). If this simplifying assumption is made, existential and disjunctive incomplete knowledge collapse; one can be defined in terms of the other. If all objects are named, the fact that there exists an object that satisfies a particular property can be expressed as the disjunction of that property applied to all the objects in the domain. However, we cannot make this sim-

plifying assumption because we are dealing with an environment where not all objects are known at plan time. Thus we not only need to study the use of disjunctive plans, we also need plans with run-time variables.

4.2 Tacit actions in conditional plans

We are going to analyze commands issued by the player that involve the execution of binary sensing actions that result in conditional plans.

In order to motivate conditional plans, let us consider an example. Suppose that the player is in a room with a locked door. She is looking around searching for a way to open the door, when the game says that there are two keys (one silver and one golden) lying on a table in front of her. Then she inputs the command "Open the door". Correctly executing this command in the state of the game described amounts to executing the following conditional plan. The plan involves taking both keys, trying the silver one in the door, and (if it fits) unlocking and opening the door; otherwise the golden key is used.

```
<init>
  take(silver_key,table)
  take(golden_key,table)
  trykey(silver_key,door)
  <branch,fits_in(silver_key,door)>
  <k+>:
     unlock(door,silver_key)
     open(door)
  <k->:
     unlock(door,golden_key)
     open(door)
```

But should the game execute this plan for the player? There is no definitive answer to this question unless we refine it further; we need to consider what the goal of such a text-adventure game is. For a start, it certainly shares a number of similarities with task-oriented dialogue systems (such as (Ferguson et al., 1996)). In particular, like task-oriented dialogue systems, our text-adventure has knowledge of the task; it models the steps involved in the task and how to talk about them. But task-oriented dialogue systems typically strive to solve the task as efficiently as possible, even if this leads to unnatural dialogue. On the other hand, for games (and indeed for tutoring systems too) efficiency in task performance and brevity is not necessarily an advantage; the longer the interaction the greater the opportunity of having a useful interactive experience (and more opportunity for learning). If we take this perspective, then a natural answer to our question would be: the game should *not* open the door for the player; rather it must force the player to perform all the steps on her own so that she will learn the task.

However, in order for a game to be engaging it cannot force the player to do repetitive tasks over and over again. In games, rules are not stated in advance; games require the skills of rule discovery through observation, trial and error, and hypothesis testing. Figuring out the rules governing the behavior of a dynamic representation is basically the cognitive process of inductive discovery, and this is challenging and motivating. But once a rule is learned, the player will no longer find it motivating to automatically apply it again and again. How to best use facts and rules that the player learned is an issue that needs to be carefully considered when deciding how a system should behave.

So, what's the answer to our question? What should the game do? Or in more general terms, when can this command (which gives rise to particular implicatures) felicitously occur? This depends on what has already happened in the game. Has the player already been through enough experiences to have the knowledge that is necessary in order to "open the door"? If yes, don't force the player to repeat the boring steps.

But how can we represent the knowledge that is necessary in order to find the conditional plan involved by this command, in order to leave the necessary actions tacit? To illustrate our explanation, let us go back to the concrete input "Open the door" and its conditional plan and analyze how it is handled by the system. The sensing action involved in the conditional plan is `trykey` defined in PKS as follows:

```
<action name="trykey">
  <params>?x, ?y</params>
  <preconds>
    Kf(accessible(?x)) ^
    Kf(locked(?x)) ^
    Kf(key(?y)) ^
    Kf(inventory_object(?y))
  </preconds>
  <effects>
    add(Kw, fits_in(?y,?x));
  </effects>
  </action>
```

Intuitively, after executing the action `trykey(?x,?y)` the agent *knows whether* a particular key `?x` fits in a locked object `?y` or not. Is this knowledge enough to find the conditional plan above? No, because it could be the case that none of the two keys fit into the door. If this is a possibility, then the conditional plan may not

achieve the goal `Kf(open(door))`. In order to rule out this possibility the following facts have to the added to the initial state of the planning problem:

```
add(Kx, fits_in(k1,c1)|fits_in(k2,c1))
```

Given this information, PKS is able to come up with the conditional plan above.

In its current version, PKS only returns disjunctive plans that will always be successful given the specification of the planning problem. It doesn't matter what the actual configuration of the world is, PKS guaranties that there will be a branch in the plan that achieves the goal. If this cannot be achieved then PKS will say that there is no plan. However, it might be the case that there is some conditional plan that is successful for most but not all configurations of the world. It would be interesting to have a planner that could provide plans for these cases, even when some of the branches will not achieve the goal.

Implementation details

Conditional plans are executed by decomposing them in disjunctive plans. For example, the conditional plan shown above can be decomposed in two disjunctive plans, namely:

```
take(silver_key,table)
take(golden_key,table)
unlock(door,silver_key)
open(door)
```

and

```
take(silver_key,table)
take(golden_key,table)
unlock(door,golden_key)
open(door)
```

These two disjunctive plans can be directly inserted in the game flow. In the game, the semantic representation of a command is in disjunctive normal form (that is, it is a disjunction of conjunction of actions). Each disjunct corresponds to a different reading of the command, hence a command's semantic representation will contain more than one disjunct if the command is ambiguous. Here, each branch of the plan can be reinserted into the game flow as a disjunct in the semantic representation of the command. Only one of the branches will be successfully executed since the sensed information is known to be exclusive (only one of the keys fits).

4.3 Run-time variables in tacit actions

In this section we are going to analyze commands issued by the player that involve the execution of existential sensing actions. Existential sensing actions result in parametric plans, that is, plans that include actions with run-time variables, values that will only be known at run time.

In order to motivate parametric plans, let us consider an example in a multiplayer game scenario. There is a player called Beatrix who has found a room with a panel where the location of all other players can be checked. Beatrix knows that in this game scenario, a player can drive herself to any other location if she knows the destination, and that in order to kill someone you have to be in the same place. Beatrix wants Bill dead and so she utters the command "Kill Bill". How do we have to represent this information so that the planner will be able to come up with a successful plan? The goal of the command can be represented with `Kf(dead(bill))` and the information about how the game world works that is already available to Beatrix can be represented with the following action schemas:

```
<action name="checklocation">
  <params>?x</params>
  <preconds>
    Kf(player(?x))
  </preconds>
  <effects>
    add(Kv, haslocation(?x));
  </effects>
</action>
<action name="drive">
  <params>?x,?y</params>
  <preconds>
    Kf(player(?x)) ^
    Kv(?y) ^
    Kf(haslocation(?x)!=?y)
  </preconds>
  <effects>
    add(Kf, haslocation(?x)=?y);
  </effects>
</action>
<action name="kill">
  <params>?x, ?y</params>
  <preconds>
    Kf(player(?x)) ^
    Kf(player(?y)) ^
    Kf(haslocation(?x)=haslocation(?y))
  </preconds>
  <effects>
    add(Kf, dead(?y));
  </effects>
</action>
```

With this information and a factual representation of the initial state the planner should return the following parametric plan. The plan involves checking Bill's location in the panel, driving to that location and killing Bill. The plan is not fully instantiated, as the actual location of Bill will only become known when the command is executed.

```
checklocation(bill)
drive(beatrix,haslocation(bill))
kill(beatrix,bill)
```

When the action `drive` is actually executed in the game, Bill's location can be obtained from the player knowledge base because the action `checklocation` will already have been executed.

5 Conclusions

In this paper we studied Lewis's broad notion of accommodation as a natural schema for treating tacit dialogue acts, tacit physical acts and tacit sensing acts in a uniform way, and we analyzed examples of the interaction among these three types of acts. In particular, we looked at sensing acts and two widely studied dialogue acts: grounding and referring. We also investigated phenomena usually studied in collaborative models of reasoning, such as exogenous events, using this same schema. Following this agenda, our final aim is to reconcile linguistic reasoning and collaborative reasoning in situated conversation.

We then turn to the question of how to model these interactions. For this purpose we integrated the planner PKS in a text-adventure game. PKS is a knowledge-based planner that is able to construct conditional and parametric plans. Such planning abilities allow the game to infer physical and sensing actions left tacit by the player. We believe that the non-traditional setup offered by the game is particularly suited to the study of the differences between collaborative task solving and other (less collaborative) types of interaction.

The work presented in this paper is in its early stages, and it is crucial to carry out an empirical test of our claims. But we believe that we have started to define a path which is worth following for two main reasons. On the theoretical side, we believe that the use of different kinds of tacit actions is omnipresent in human interaction and will help generalize the theory of accommodation. On the practical side, sensing actions are an essential component if we want to build situated dialogue systems that are able to interact in a realistic way.

References

F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel. 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

D. Beaver and H. Zeevat. 2007. Accommodation. In *The Oxford Handbook of Linguistic Interfaces*, pages 503–539. Oxford University Press.

D. Beaver. 1994. Accommodating topics. In *The Proceedings of the IBM and Journal of Semantics Conference on Focus*.

L. Benotti. 2007. Incomplete knowledge and tacit action: Enlightened update in a dialogue game. In *Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–24, Rovereto, Italy.

L. Benotti. 2008. Accommodation through tacit dialogue acts. In *Conference on Semantics and Modelisation*, Toulouse, France.

D. DeVault and M. Stone. 2006. Scorekeeping in an uncertain language game. In *The 10th Workshop on the Semantics and Pragmatics of Dialogue*, University of Potsdam, Germany.

O. Etzioni, S. Hanks, D. Weld, D. Draper, N. Lesh, and M. Williamson. 1992. An approach to planning with incomplete information. In *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning*, pages 115–125.

G. Ferguson, J. Allen, and B. Miller. 1996. TRAINS-95: Towards a mixed-initiative planning assistant. In *Proceedings of the Conference on Artificial Intelligence Planning Systems*, pages 70–77, Edinburgh, Scotland.

R. Fikes, P. Hart, and N. Nilsson. 1972. Learning and executing generalized robot plans. *Artificial Intelligence*, 3:251–288.

J. Hobbs, M. Stickel, D. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

H. Kautz and B. Selman. 1999. Unifying SAT-based and graph-based planning. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 318–325, Stockholm, Sweden.

J. Kreutel and C. Matheson. 2003. Context-dependent interpretation and implicit dialogue acts. In *Perspectives on Dialogue in the New Millenium*, pages 179–192.

D. Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–359.

E. Martinez and Y. Lesperance. 2004. Web service composition as a planning task: Experiments using knowledge-based planning. In *Proceedings of the Workshop on Planning and Scheduling for Web and Grid Services*, pages 62–69.

R. Petrick and F. Bacchus. 2002. A knowledge-based approach to planning with incomplete information and sensing. In *Proceedings of the Sixth International Conference on Artificial Intelligence Planning and Scheduling*, pages 212–221.

R. Stalnaker. 1998. On the representation of context. *Journal of Logic, Language and Information*, 7(1):3–19.

R. Thomason, M. Stone, and D. DeVault. 2006. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In *Presupposition Accommodation*. Ohio State Pragmatics Initiative.

What eye believe that you can see: Conversation, gaze coordination and visual common ground (DSiJA)

Daniel C. Richardson
University of Reading
dcr@eyethink.org

Rick Dale
University of Memphis
radale@memphis.edu

John M. Tomlinson Jr
UC Santa Cruz
otomlins@ucsc.edu

Herbert H. Clark
Stanford University
herb@psych.stanford.edu

Abstract

We can only share information because of how much we share already. Conversation is supported by the common ground between us, such as the beliefs and the visual context that we have in common. It has been shown that both of these components determine how we communicate, yet it is not clear how they interact. In a new paradigm, we separated the fact that a visual scene was shared or not and the *belief* that a visual scene was shared or not. We quantified the effects of these factors upon joint attention by measuring the coordination between conversants' eye movements. Participants had a conversation about a controversial topic, such as the Iraq war. The discussion was first framed by four short videos of actors espousing tendentious views. Participants discussed their own views while they looked at either a blank screen or pictures of the four actors. Each believed (correctly or not) that their partner was either looking at a blank screen or the same images. We found that both the presence of the visual scene and beliefs about its presence for another influenced participants' discussion and the coordination of their joint attention.

Introduction

"Can you pass me the thingy for the whatsit?" Penny asked John. It is hard to imagine a more vacuous sentence. Yet to John, it was a precise instruction. He replied, "Not too much". The content of this communication came not from the words spoken as much as the rich body of knowledge that John and Penny shared. This is termed their common ground (Clark, 1996). A conversation that morning (about the guests coming to dinner), specific knowledge that they shared (concerning one guest's tastes) and their current visual context (Penny standing in front of the stove and John in front of a particular drawer) restricted her reference to a tool that was within John's reach that would allow Penny to crush some garlic into a casserole, although not too much.

In this paper, we examine one aspect of common ground: the shared visual context. What role does this information play in the production and comprehension of spontaneous dialog? Would Penny have spoken the same oblique phrase, and would John have understood it, if he had been facing the other way? Clark and Marshall (1981) argued that we interpret ambiguous references using the co-presence heuristic. Only items seen by both conversants are considered as possible referents. This claim has been tested in various 'reference game' studies with mixed results. A speaker refers to an object which is in one or both of the participants' sight, and as a consequence there are sometimes changes in the listener's eye movements (Hanna, Tanenhaus & Trueswell, 2003) and the speaker's manner of reference (Haywood, Pickering & Branigan, 2005), and sometimes not (Keysar, Barr & Brauner, 2000; Horton & Keysar, 1996). Language does more than lead people to objects, however. It can describe people, ideas and opinions that are abstract or simply absent (Spivey & Richardson, in press). In contrast to reference game studies that have a speaker, a listener and a reference to an object, our experiment examined the role of visual context when two people have an extended conversation. The situation was analogous to John and Penny discussing the political views of their dinner guests for that evening, while looking at the meal they are to serve. Objects in the shared visual scene were not the content the utterances, but, in a more germane sense, provided a visual context for the discussion. Whether or not conversants chose to incorporate this shared visual information becomes a more interesting question, since the constraints of a referencing task do not demand that they do so.

There are two ways that visual context could play a part in a conversation (Keysar 1997). First, it provides information. The sight of certain objects or people may relieve the burden of memory or lexical access for a speaker, and help disambiguate language sounds or structure for a listener. Certainly, in speech production (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998) and comprehension

(Cooper, 1974; Kamide, Altmann, & Haywood, 2003; Knoeferle & Crocker, 2006; Richardson & Matlock, 2007; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) eye movements to a visual scene are closely linked, moment by moment, to linguistic processes. Second, if a conversant believes that her conversational partner can see certain things, then she can interpret utterances in relation to the common ground, and make remarks relying on the fact that this information will be available to help her listener. Beliefs about shared knowledge influence speech. Speakers change their descriptions of locations in New York (Isaacs & Clark, 1987) or famous faces (Fussell & Kraus, 1992) depending upon their estimation of how much relevant knowledge the listener might have. But what of beliefs about shared visual context?

All reference game studies of visual common ground and conversation have confounded the belief that visual information is shared with the fact of it being shared. Our experiment separated these factors for the first time. A visual image was either present or absent for both conversants, and both believed that it was either present or absent for their partner. We hypothesized that both the presence of the visual context for an individual and the belief in its presence for another would influence their conversation. We captured the success of this joint activity by measuring the coordination between conversants' eye movements as they talked and looked at a shared display.

Gaze coordination and conversation

The temporal dynamics of gaze coordination are intertwined with discourse. In the first quantification of gaze coordination, Richardson and Dale (2005) recorded the eye movements of speakers talking spontaneously about a TV show while looking at pictures of its cast members. These speeches were played back to listeners who were looking at the same display. Cross-recurrence analysis (Zbilut, Giuliani, & Webber, 1998), measured the degree to which speaker and listener's eye positions overlapped at successive time lags (see below for an explanation). From the moment a speaker looked at a picture, and for the following six seconds, a listener was more likely than chance to be looking at that same picture. The overlap between speaker and listener eye movements peaked at about 2000ms. In other words, two seconds after the speaker looked at a cast member, the listener was most likely to be looking at there too. The same eye movement coupling when two participants had a live spontaneous dialog and looked at the same images (Richardson, Dale & Kirkham, 2007). On this occasion, gaze recurrence peaked at 0ms, presumably representing the average of each conversant acting as speaker and then listener. This coordination was achieved in virtue of the knowledge conversants shared. Gaze recurrence was increased when conversants heard the same (rather than different) encyclopedia passages about Salvador Dali prior to discussing one of his paintings. Closer gaze recurrence appears to facilitate communication. When pictures in a display flashed in time with the speakers' fixations, it caused listeners' eye movements to follow the speakers' more closely. Consequently, listeners answered comprehension questions faster than those who had seen a randomized sequence of flashes (Richardson & Dale, 2005). Since gaze recurrence is causally connected to what conversants know and remember, we predicted that it would reveal effects of what they see and what they believe each other can see.

Visual context for self and for others

Our conversants watched four actors give their views on a contentious topic and then discussed the topic between themselves. We manipulated two factors. First, the actors could either be present on the screen for each conversant, or absent, replaced by an empty two by two grid. We termed this as the visual context *for self*. Second, each participant was told prior to the discussion that the actors were either present or absent on the screen of their conversational partner. We termed this the visual context believed *for other*. We refer to the combinations of these conditions by stating the visual context *for self* followed by believed *for other*. For example, *present - absent* refers to the condition in which both participants could see a visual scene, but believed that their partner could not. To be clear: the two conversants were in a symmetrical situation, always looking at the same thing as each other and always believing the same thing as each other.

Conversants had beliefs that were factually incorrect in the present-absent and absent-present conditions. We inculcated these beliefs by a slight deception. At the start of each conversation, both participants read the words

'You are participant B'. They were then shown, for example, a blank screen and told, 'Participant A is looking at a blank screen. Participant B should still be looking at the pictures. Please say, 'yes' if this is true'. Both participants, seeing a blank screen and believing themselves to be participant B, said 'yes'. They also heard each other saying yes, and interpreted this as their partner, participant A, confirming that looked at a blank screen.

Our first, straightforward prediction was that the presence of a visual context for self would increase gaze recurrence, since during speech production and comprehension relevant visual objects are fixated. We did expect some recurrence between gaze patterns even when the screen was empty, however. During language comprehension and memory tasks, empty locations of a screen can be systematically fixated when a reference is made to items or events that were previously there (Altmann, 2004; Hoover & Richardson, in press; Richardson & Spivey, 2000; Richardson & Kirkham, 2004; Spivey & Geng, 2001). The more contentious question is what will be the effect of the visual context that is believed for others. We put forward three possibilities: (1) visual context is ignored, and so there will be no effect of beliefs about it, (2) visual context is exploited, and so the belief that more of it is shared will increase gaze coordination (3) visual context is compensated for, and so the belief that it differs between conversants will increase gaze coordination. These possibilities are not exhaustive, but there is support for each in the literature.

(1) Ignoring visual context. Listeners can seem strikingly egocentric. They ignore, in the first instance at least, the fact that a speaker's visual perspective differs from their own (Keysar, Barr & Horton, 1998). In a reference game, an array of objects was placed between the participant and a confederate (Keysar, et al, 2000). Some of the objects were occluded so that only the participant could see them. For example, the participant could see three candles of different sizes, but the confederate could only see the larger two. When the participants were asked for 'the smallest candle', they were more likely to look at the very smallest candle. Since it could not be seen by the confederate, it could not have been the intended referent. Therefore, mutual knowledge is a non-existent or partial constraint upon speech comprehension. In our case, conversants are not even directing each other to pick up objects but discussing current affairs. The prediction from these results is that when we manipulate the visual context that is believed for others it will have no effect on behaviour.

(2) Exploiting visual context. Subsequent work has suggested that the 'partial constraint' of mutual knowledge can become dominant with slightly different participants or circumstances. Native speakers of Mandarin come from a culture that has a greater focus on other people during social interactions (Markus & Kitayama, 1991). In the same reference game, they almost never failed to take into account the speaker's visual perspective (Wu & Keysar, 2007a). However, even English speaking participants are not insensitive to common ground constraints. When two possible candidates for 'the smallest candle' were on display, listeners immediately fixated the candle that was in the visual common ground, and ignored the one that was blocked from the speaker's view (Hanna, Tanenhaus & Trueswell, 2003). When a speaker begins to ask a question about an object, a listener is more likely to fixate those that are hidden from the speaker's view (Brown-Schmidt, Gunlogson & Tanenhaus, in press). Speakers will use names for objects that they believe to be known to the listener (Isaacs & Clark, 1987; Metzing & Brennan, 2003), though they sometimes overestimate the degree that the knowledge is shared (Wu & Keysar, 2007b). These results suggest that common ground will be used if it is available. This hypothesis predicts a main effect of the belief condition: whether or not the visual context is present *for self*, gaze coordination will increase if it is believed to be present *for others*.

(3) Compensating for visual context. Our third hypothesis is that there will be an interaction between the visual context for self and the visual context that is believed for others. When conversants believe there is a difference between what they see and what their partners can see, they will seek to redress the imbalance. Speakers use more gestures when they are describing a toy that listeners have not played with before (Gerwing & Bavelas, 2004) or a location within a picture that they have not seen (Holler & Stevens, 2007). Speakers produce better explanations when they believe their listeners do not have access to a diagram

(Bromme, Jucks & Runde 2005), and provide disambiguating information when it is not present in the visual common ground (Haywood, Pickering & Branigan, 2005). This suggests that when conversants believe there is a mismatch between their visual context and their partners' in our task, they will boost their efforts to establish common ground, leading to better gaze coordination. This hypothesis predicts that gaze coordination will be highest in the *present-absent* and *absent-present* conditions.

Methods

Participants

112 undergraduates from the University of California, Santa Cruz took part in exchange for course credit. The data from 19 participants was discarded due to failures in calibration, resulting in 37 dyads with two usable data series.

Apparatus

Each participant sat in a cubicle in a reclining chair, looking up at an arm-mounted 19" LCD 60cm away with a *Bobax3000* remote eye tracker mounted at the base. They wore a headset with a boom mic. The experimenter controlled when the participants could hear the stimuli, each other's voices, or the experimenters voice. For each participant, an iMac calculated gaze position approximately 30 times a second, presented stimuli and recorded data. A third Apple Mac computer synchronized the trials and data streams from the iMacs and saved an audio-video record of what was seen, heard and said during the experiment, superimposed with gaze positions.

Design

Four different opinion pieces were written for each of eight contentious topics (the iraq war, vegetarianism, drugs in sport, UCSC professors, UCSC campus expansion, violence in video games, online social networks, and gay marriage). The opinion pieces were delivered straight to camera by sixty four different actors, producing movies that varied from 8 to 20 seconds in length. For each pair of participants, the topics were randomly allocated across the four experimental conditions.

Procedure

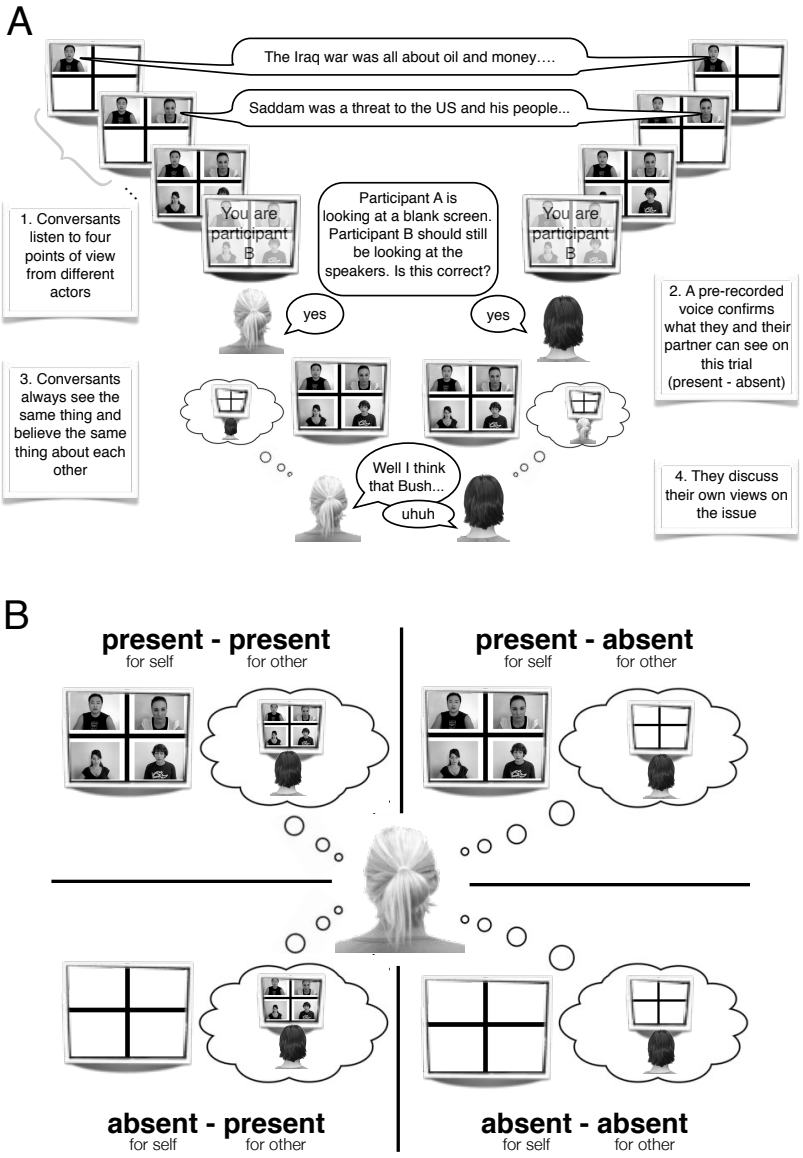
Participants were introduced to each other in the laboratory waiting room. They then sat in adjacent cubicles and underwent a brief calibration routine of roughly a minute. The trial design is shown in Figure 1A. First the movies were shown, one at a time, in each of the quadrants of a 2 x 2 grid. Location and order of presentation were randomized, but were identical for each participant. Each movie ended with a freeze frame which remained on screen. In the two *absent for self* conditions, pictures of the four actors faded from view at the end of presentations, leaving an empty grid. In the *present for self* conditions, the pictures remained in view. The words 'You are participant B' appeared on the screen *for both participants*. The participants then heard a prerecorded voice saying, "Please discuss these issues," and the experimenter activated the audio link between cubicles.

Figure 1B represents the different experimental manipulations that were introduced at this stage of the trial. In the *absent-absent* and *present - present* condition, participants heard "You should both now [be looking at a blank screen / be able to still see the speakers on screen]". In the *absent-present* and *present-absent* conditions they heard "Participant A should [be looking at a blank screen / still be able to see the speakers]. Participant B should [still be able to see the speakers / be looking at a blank screen]". Across all conditions, they were then asked, "Please say 'yes' if this is the case". Once they had affirmed, the experimenter initiated the conversation. Participants typically talked for between one and three minutes before the experimenter decided that the topic had been exhausted, and the trial was terminated.

Data analysis

We quantified the gaze coordination between conversants by generating categorical cross-recurrence plots. This technique depicts the temporal structure between time series (Zbilut, Giuliani, & Webber, 1998), and has been used to capture the subtle entrainment of body sway during conversation (Shockley, Santana & Fowler, 2003) and the interrelationships between a child and care givers' language use (Dale & Spivey, 2006). In our case, points of recurrence are defined as the times at which both conversants are fixating the same screen quadrant. For each trial, we took the first minute of eye movement data, added up all the points of recurrence and then divided by the total number of possible points to get a recurrence percentage (for a detailed explanation, see Richardson & Dale, 2005). Here, the possible points of recurrence were defined as the times at which at least one of the conversants had their

eyes on the screen. The next step was to lag one of the data streams by 10ms, so that 0ms on one data stream was aligned with 10ms on the other. Again, all the points of recurrence were calculated. This represents the degree to which one conversant is looking at the same thing as the other conversant 10ms later. Since in our experiment the participants had symmetrical roles, we averaged the recurrence for a lag of 10ms and -10ms. A full cross-recurrence analysis consisted of calculating the recurrence for all possible alignments, or lag times, of the two data series.



Differences in gaze coordination would be produced by differences in how conversants spoke to each other. Therefore, our gaze analysis was supported by a secondary analysis of the speech of a subset of ten pairs of participants. We coded all eight conversations for each dyad and focused on utterances which made any reference at all to the views or appearance of the actors, as gaze coordination has been found to increase during moments of direct reference to a visual scene (Richardson & Dale, 2005). We counted the number of such references in each conversation and analysed three properties. Firstly, each reference was termed either 'visual' if it made any mention of visual properties of the actor, or 'factual' if it contained only (non-visual) information relating the actors' viewpoint. Secondly, all references that occurred at the end of phrases were coded for intonational contours called *boundary tones*. Of interest were *rising boundary tones*, in which the intonation goes up in pitch at the end of a sentence from either the lower or higher part of a speaker's pitch range (Pierrehumbert & Hirschberg, 1990). They are thought to serve as implicit requests for confirmation from the listen, and hence play a role in establishing information in the common ground (Allen, 1984; Fletcher, Stirling, Mushin & Wales, 2002). Lastly, we calculated the proportion of speakers' references that elicited a back-channel response from the the listener such as *ok*, *uh-huh*, *right*. Back-channels had to be directly following or overlapping the speakers' reference and consist of one word. These responses are thought to serve the function of coordinating dialogue and signaling to the speaker that she has been understood (Schegloff, Jefferson & Sacks, 1977), and hence, like the rising contours, their frequency serves as a indicator of conversants' efforts to establish common ground.

Results

Gaze coordination was increased by the presence of a visual context for the self, and modulated by a belief in the presence of a visual context for the other. When the visual context was absent for conversants, recurrence was higher if they believed their partners could see it. Conversely, when the visual context was present, recurrence was higher if they believed that that their partners could not. Figure 2A shows how gaze recurrence changed across conditions at different time lags. With the exception of the *absent-absent* condition, recurrence peaked at or around 0ms and trailed off as the distance between conversants' gaze patterns increased, replicating the pattern produced by spontaneous conversation that was observed by Richardson, Dale & Kirkham (2007). Following their analysis, differences between conditions were analyzed by averaging recurrence within a window of +/- 3000ms, to capture the typical periods in which both conversants were acting as speakers and listeners. A 2 (*for self: present/absent)* x 2 *(believed for other: present/absent)* ANOVA showed an significant interaction between theses effect ($F(1,36)=4.5$, $p<.05$), as well as the expected main effect of visual context ($F(1,36)=12.4$, $p<.001$). Recurrence in each of the conditions was compared to randomized baselines. Within the critical window, recurrence was higher than chance for the present-present ($t(36)=3.3$, $p<.001$), present-absent ($t(36)=3.7$, $p<.001$) and absent-present ($t(36)=2.8$, $p<.01$) conditions, but not for the absent-absent ($t(36)=1.6$).

Conversants' speech was also influenced by an interaction between the presence of a visual context for the self and a belief in the presence of a visual context for the other (see figure 2B). Whilst conversants simply made more references overall to the actors when they were on the screen in front of them, our other measures revealed interactions between conditions. In the *present-absent* and *absent-present* conditions, references were more likely to employ factual (non-visual) properties (for example, 'the guy who said that Iraq was about oil' rather than 'the guy in the red shirt who was against the war'). In these conditions, references that came at the end of phrases were more likely to end in a rising contour (as an implicit request for confirmation), and references were more likely to receive a back-channel response from the listener. Our gaze analyses show that these efforts to boost common ground in those conditions did indeed correspond to an increase in gaze recurrence.

Our measures of conversants' references were analysed by a 2 (*for self: present/absent)* x 2 *(believed for other: present/absent)* ANOVAs. There was a main effect of the presence of the visual context *for self* ($F(1,9)=8.3$, $p<.05$) on the number of references made. There were significant interactions between the visual context *for self* and *believed for other* on the proportion of references that mentioned factual properties ($F(1,9)=6.8$, $p<.05$), the proportion of references using rising contours ($F(1,9)=5.3$, $p<.05$), and the proportion of references that received back channel responses ($F(1,9)=6.35$, $p<.05$). No other main effects or interactions were significant across our four measures.

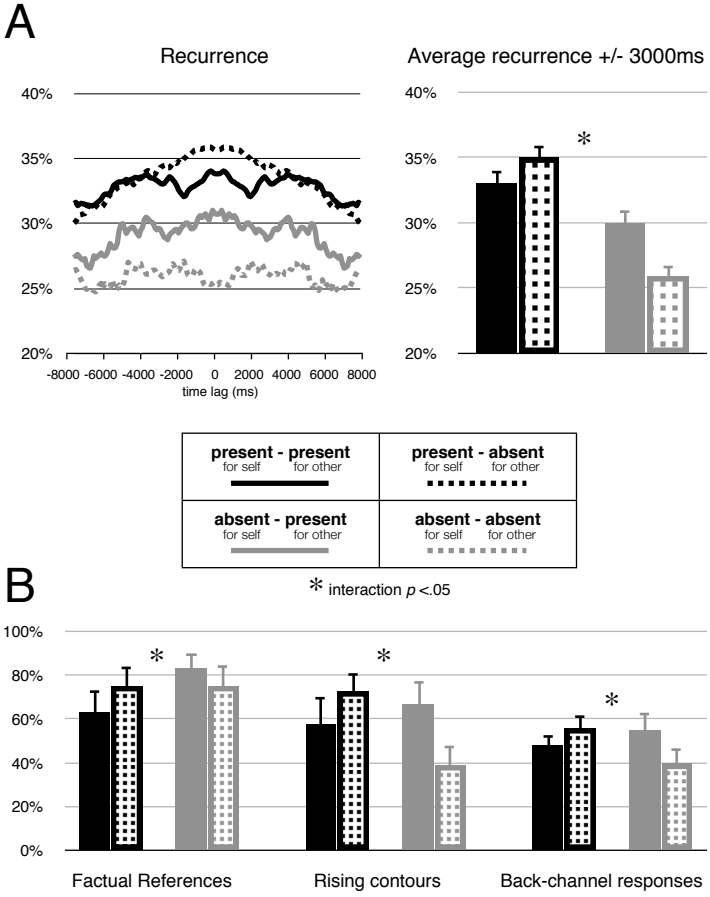


Figure 2. Results of (a) gaze analysis and (b) speech analysis

Conclusion

Reference game studies have given a mixed view of the role of visual common ground in conversation. Sometimes the visual perspective of a speaker has little effect on a listener (Keysar et al 2000), and sometimes it has an immediate constraint on an ambiguous reference to an object (Hanna et al, 2003). For a number of reasons, our experiment might have found no influence of visual common ground. The conversations were not about the actors on display, but concerned politics, sports and campus life. The effect of believing that a conversational partner could see the actors was deconfounded from the fact of them being seen by each conversant. Lastly, the particular circumstances of who saw or was believed to see what changed on a trial by trial basis, demanding that conversants keep track of shifting visual common ground constraints.

Under these conditions, conversants could have ignored the whole issue of what they believed each other could see, but they did not. They could have employed a quick and expedient technique of exploiting visual information when it was believed to be present, and ignoring it otherwise. Instead, they maintained an awareness of what each other could see on each particular trial, when they believed there to be an imbalance between their own view and their partners' they sought to compensate for that difference. They tended to refer to actors' viewpoints via a non-visual route, ask for confirmation that their messages had been understood, and signal understanding to each other via back channel responses. As a result, when they looked at the pictures their eye movements were more tightly coordinated if they believed they were talking to someone who was looking at nothing. Conversely, while looking at an empty screen their eye movements were more closely coordinated if they believed that each other could still the actors.

Coordinating joint attention is essential for successful communication (Clark, 1996; Clark & Brennan, 1991; Schober, 1993). It may even be the basis for pre-linguistic learning (Baldwin, 1995). In spontaneous conversation, people are able to couple their gaze despite ambiguities and disfluencies in the speech stream. This remarkable coordination is achieved in virtue of the background knowledge they share (Richardson, et al., 2007), their sensitivity to each others' pragmatic constraints (Hanna & Tanenhaus, 2004) and moment-by-moment domains of reference (Brown-Schmidt & Tanenhaus, in press). Here we have shown that conversants are also attuned to both their own visual context and what they believe each other can see. They will even coordinate their gaze around an empty screen in the mistaken belief that each other can see something. It is the net effect of these multiple constraints that restrict the referent of "the thingy for the whatsit" from a universe of objects to a garlic press.

References

Allen, K. (1984). The high rise terminal contour. *Australian Journal of Linguistics*, *4*, 19-32.

Altmann, G.T.M. (2004) Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition*. 93, 79-87.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), *Joint attention: its origins and role in development*. Hillsdale, NJ: Lawrence Erlbaum.

Bromme, R., Jucks, R. & Runde, A. (2005). Barriers and biases in computer-mediated expert–layperson communication: An overview and insights into the field of medical advice. In R. Bromme, F.W. Hesse & H. Spada (Eds.). *Barriers and biases in computer-mediated knowledge communication—and how they maybe overcome* (pp. 89–118). New York: Springer.

Brown-Schmidt, S. & Tanenhaus, M.K. (in press). "Real-time investigation of referential domains in unscripted conversation: a targeted language game approach," *Cognitive Science*.

Brown-Schmidt, S., Gunlogson, C. & Tanenhaus, M. K. (in press). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*.

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

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

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge: Cambridge University Press.

Dale, R. & Spivey, M. J. (2005). Categorical recurrence analysis of child language. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 530-535). Mahwah, NJ: Lawrence Erlbaum.

Dale, R. & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and

caregivers in conversation. *Language Learning*, 56, 391-430.
Fletcher, J., Stirling, L., Mushin, I., & Wales, R. (2002). Intonational rises and dialog acts in the Austrialian English Map Task. Language and Speech, 45(3), 229-252.
Gerwing, J., & Bavelas, J. B. (2004). Linguistic influences on gesture's form. *Gesture*, 4, 157–195.
Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274-279.
Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory & Language*, 49(1), 43-61.
Haywood, S.L., Pickering, M.J., & Branigan, H.P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science,* 16, 362-366.
Hoover, M. A. & Richardson, D. C. (in press). When Facts Go Down the Rabbit Hole: Contrasting Features and Objecthood as Indexes to Memory. *Cognition*
Holler, J. & Stevens, R. (2007) The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology,* 26, 1-25.
Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59, 91-117.
Isaacs, E. A., & Clark, H. H. (1987). References in conversations between experts and novices. *Journal of Experimental Psychology: General* , 116, 26-37
Keysar, B. (1997). Unconfounding common ground. *Discourse Processes*, 24, 253-270.
Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Sciences*, 7, 46-50.
Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science,* 11(1), 32-38.
Knoeferle, P. & Crocker, M (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, 30, 481-529.
Markus, H., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review,* 98, 224–253
Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory & Language*, 49(2), 201-213.
Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66(2), B25-B33.
Pierrehumbert, J. & Hirshberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgen, & M. E. Pollack (Eds.), *Intentions in Communications*. (pp. 271-311). Cambridge, Mass.: The MIT Press.
Richardson, D.C & Dale, R. (2005). Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and its Relationship to Discourse Comprehension. *Cognitive Science,* 29, 1045–1060.
Richardson, D.C., Dale, R. & Kirkham, N.Z. (2007) The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science*, 18 (5), 407-413
Richardson, D. C. & Kirkham, N.Z. (2004). Multi-modal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General*, 133 (1), 46-62.
Richardson, D.C. & Matlock, T. (2007) The integration of figurative language and static depictions: an eye movement study of fictive motion, *Cognition*, 102, 129-138
Richardson, D. C. & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, 76, 269-295.
Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-repair in the organization of repair in conversation. *Language*, 53, 361-382.
Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, 47(1), 1-24.
Shockley, K., Santana, M.V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception & Performance*, 29(2), 326-332.
Spivey, M. J., & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychological Research,* 65, 235-241.
Spivey, M. & Richardson, D. (in press). Language embedded in the environment. In P. Robbins and M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge, UK: Cambridge University Press.
Tanenhaus, M. K., Spivey Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632-1634.
Wu, S. & Keysar, B. (2007a). Cultural effects on perspective taking. *Psychological Science,* 18, 600-606.
Wu, S., & Keysar, B. (2007b). The effect of communication overlap on communication effectiveness. *Cognitive Science,* 31, 169-181.
Zbilut, J. P., Giuliani, A., & Webber, C. L., Jr. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters,* 246, 122-128.

Adapting the use of attributes to the task environment in joint action: results and a model

Markus Guhe
University of Edinburgh
Linguistics and English Language & HCRC
40 George Square
Edinburgh, EH8 9LL
m.guhe@ed.ac.uk

Ellen Gurman Bard
University of Edinburgh
Linguistics and English Language & HCRC
40 George Square
Edinburgh, EH8 9LL
ellen@ling.ed.ac.uk

Abstract

Speakers use referring expressions to identify an object in the environment. To generate a referring expression, features of the intended referent have to be selected that distinguish the object from the other potential referents. Current accounts of referring expressions consider a number of factors that influence the choice of features but ignore the influences of the task environment. In particular, they do not address how these influences change the generation of referring expressions over an extended period of time. We present results of how colour terms are used to describe landmarks in a task oriented dialogue (a route communication task) and describe a computational cognitive model of the observed adaptations over time.

1 Introduction

Much attention in recent computational as well as psychological research on language has been given to the linguistic problem of the use and generation of referring expressions. Referring expressions are linguistic expressions that identify either a referent entity in the real world or a discourse entity in the form of an antecedent. Referring expressions serve the purpose of distinguishing the target or referent from the set of other possible referents in the given context, called the distractor set. For example, in the set of objects in Figure 1, *the black cup* and *the small, black cup* would both succeed in distinguishing the cup at the lower left (the referent) from the other two objects (the distractor set).

A speaker wanting to pick out that small, black, cup at the lower left of the array could use any of the attributes in the expressions just given. Computational approaches to generating referring expressions often produce expressions that, if possible, uniquely and minimally select the target object. But such algorithms are computationally costly and may not be helpful in modelling human behaviour: People (1) produce non-minimal expressions, which contain redundant information (e.g., Pechmann 1989) and (2) interpret such expressions more easily (e.g., Paraboni, van Deemter and Masthoff 2007).

Figure 1: A simple domain of reference: for each object, the other are distractors

A prominent account of how human-like, non-minimal referring expression can be generated is the algorithm by Dale and Reiter (1995), which by now has many extensions (see van der Sluis (2005) for a recent overview). This algorithm incrementally tests whether using an attribute in a referring expression will rule out distractor objects. The attributes are tested according to a preference list that is fixed beforehand. For the domain used in Figure 1, for example, this preference list could be <type, colour, size>. Identifying the object to the right would then produce the non-minimal expression *large, white cup* by first adding the type attribute (which has a special status and is always added), then by adding *white* (because it removes the object in the lower left from the distractor set), finally by adding *large* (because it removes the object in the top

left). Non-minimal expressions arise simply because a selected attribute is never de-selected, even if a subsequently selected attribute makes it redundant.

While these approaches deal with which of the available possibilities to describe the target object is chosen, they do not account for the adaptations that a speaker makes over time to the demands of the current task environment. The computational as well as the psycholinguistic paradigms typically lack history: On each trial a participant (or algorithm) is presented with a picture like Figure 1 and instructed to produce a suitably distinguishing expression. The trial terminates without feedback and is followed by others, presenting different objects and distinguishing features. How the fourth target is distinguished from its distractors might actually owe something to the participant's experience with the first three, and our work attempts to discover and model such effects of experience.

We examine referring expressions in an unrestricted, task-oriented dialogue in which the interlocutors get natural feedback on failures of reference and refer to many different objects. We use a variant of the HCRC Map Task (Anderson et al. 1991) in which a player who can see the route on a schematic map describes it to a fellow player who must reproduce it. Each map is populated with cartoon landmarks, distinguished by several different features. We have shown that the use of features changes across first mentions as players pursue their task (Guhe and Bard 2008). In the present paper we ask how and why the changes take place. Colour is a perceptually salient property, usually one of the first tested in the incremental Dale and Reiter type algorithms. In our experiment, however, we set unreliability against salience: Colour is an unreliable distinguisher. In contrast, each map allows for use of a reliable attribute, too, (shape, number, kind or pattern). Thus, our participants need to use the adaptive attributes but waste time and can cause misunderstandings using the unreliable one.

In this paper, we report how the use of colour terms changes over the course of the experiment and present a simple computational cognitive model of this change. More precisely, we describe how the utility of the colour feature influences the Instruction Giver's choice of whether to use colour in introductory referring expressions. The model offers an explanation of this change in terms of Anderson's rational analysis (Anderson 1990; Anderson and Schooler 1991). Rational analysis is the core mechanism in ACT-R's utility-based production selection (Anderson 2007) and is a variant of utility learning mechanisms found in reinforcement learning or the delta rule (Sutton and Barto 1998). In brief, rational analysis says that human memory reflects the frequency of events in the environment, making more frequent experiences easier to retrieve and corresponding behaviours more likely to be used. By using rational analysis our model goes beyond existing accounts of use and generation of referring expressions in that it reveals the environmental influences on these processes.

2 Comparison to existing research

The problem of whether the use of features changes with the demands of the task environment has scarcely been addressed in the literature. Although Brennan and Clark's (1996) conceptual pacts address changes in referring expressions, these changes are about how speakers refer to objects after they have been introduced. However, our questions here address the overall use of features in referring expressions over the course of many interactions. To exclude effects of conceptual pacts we are only analysing the use of introductory (first) mentions of landmarks.

Garrod and Doherty (1994) describe how a community of speakers establishes a sublanguage in referring to entities. We are concerned with the internal structure of the referring expressions themselves and propose a utility-based explanation instead of one based on precedence and salience.

There is some evidence that extra-linguistic factors play a role in generating referring expressions. For example, Arnold and Griffin (2007) show that the presence of a second character influences the choice of whether to use a pronoun or the character's name for references following the introductory mention. This is true even if the characters differ in gender, so that the name does not disambiguate any more than the pronoun. Arnold and Griffin argue that the reasons for this behaviour lie in the speakers' cognitive load when they generate the referring expression.

This is part of another strand of findings in which the cooperative view on dialogue (e.g. Clark 1996) is changed towards a speaker-oriented view (e.g. Bard et al. 2000). In this view, the speaker makes the general assumption that what he/she knows is shared knowledge. Only if problems arise in the dialogue, e.g. by explicit feedback from the listener, might the speaker adapt to the listener's needs. In fact,

even if overspecified referring expressions (Dale and Reiter 1995; Paraboni, van Deemter and Masthoff 2007; Pechmann 1989) help the listener to identify the target object, the speaker also profits in terms of a generation process of greatly reduced complexity. Since both – speaker and listener – benefit from using such referring expressions, the communicative strategy cannot be attributed uniquely to concerns for the listener's needs. In our task, however, the colour feature is counterproductive in the majority of cases, because it does not match between the two maps. So the speaker's assumption about the usefulness of the salient feature colour are mistaken.

Another related line of research is the use of machine learning techniques to extract the way attributes are selected for modified versions of the Dale and Reiter algorithm (Jordan and Walker 2005). Although these algorithms already incorporate psychological findings, e.g., conceptual pacts, they only provide global adaptations to properties of linguistic corpora and do not account for changes over time and for adaptations to the properties of the task environment.

3 Experiment

3.1 Task

The experiment is a modified Map Task (Anderson et al. 1991). The Map Task is an unscripted route-communication task in which an Instruction Giver and an Instruction Follower each have a map of the same fictional location. The Giver's map contains a route that is missing on the Follower's map. The dyad's goal is to recreate the Giver's route on the Follower's map.

The dialogue partners use the landmarks on the maps to navigate from START (shared) to FINISH (only on the Instruction Giver's map).

3.2 Materials, procedure, data collection

Materials. Some landmarks differ between the two maps. In our experiment they can differ by:

1. Being absent on one of the maps or present on both;
2. Mismatching in a feature between the two maps (most notably colour);
3. Being affected by 'ink damage' that obscures the colour of some landmarks on the Instruction Follower's map.

There are four attributes which also distinguish landmarks. Each serves for two different kinds of landmarks:

1. *Number* (bugs, trees),
2. *Pattern* (fish, cars),
3. *Kind* (birds, houses/buildings),
4. *Shape* (aliens, traffic signs).

Three crossed independent variables determine the nature of Giver–Follower map pairs:

1. *Homogeneity*: whether the landmarks on a map are of just one kind (single) or of different kinds (mixed).
2. *Orderliness*: whether the ink blot on the Instruction Follower's map obscures a contiguous stretch of the route (orderly) or a non-contiguous stretch (disorderly). The number of obscured landmarks is constant.
3. *Animacy*: whether the landmarks on a map are animate or inanimate (thus, on the mixed maps there are only landmarks from the 4 inanimate or the 4 animate kinds of landmarks).

The maps in Figure 2 are a pair of Giver and Follower maps for the disorderly, mixed tree condition. Thus, the maps contain mainly trees but also other inanimate objects (mixed), and the Follower's map shows multiple, non-contiguous ink blots (disorderly).

Procedure. Participants are told that the maps are 'of the same location but drawn by different explorers'. They thus know that the maps can differ but not where or how. They are instructed to recreate the route on the Follower's map as accurately as possible.

Each dyad did 2 simple training maps and then completed a set of 8 maps, one for each kind of landmark. The maps were counterbalanced with respect to the experimental conditions. After the fourth map, the role of Instruction Giver and Instruction Follower were exchanged.

To reduce the variability of words and concepts used in the unrestricted dialogues, each participant was prompted textually to provide standard type names for a few landmarks that would occur on the following map.

Setup and data collection. Participants sat in front of individual computers, facing each other, but separated by a visual barrier.

This research is part of a larger multimodal project. The communication was recorded using 5 camcorders. The Giver was eye tracked using a remote eye tracker. Speech was recorded using a

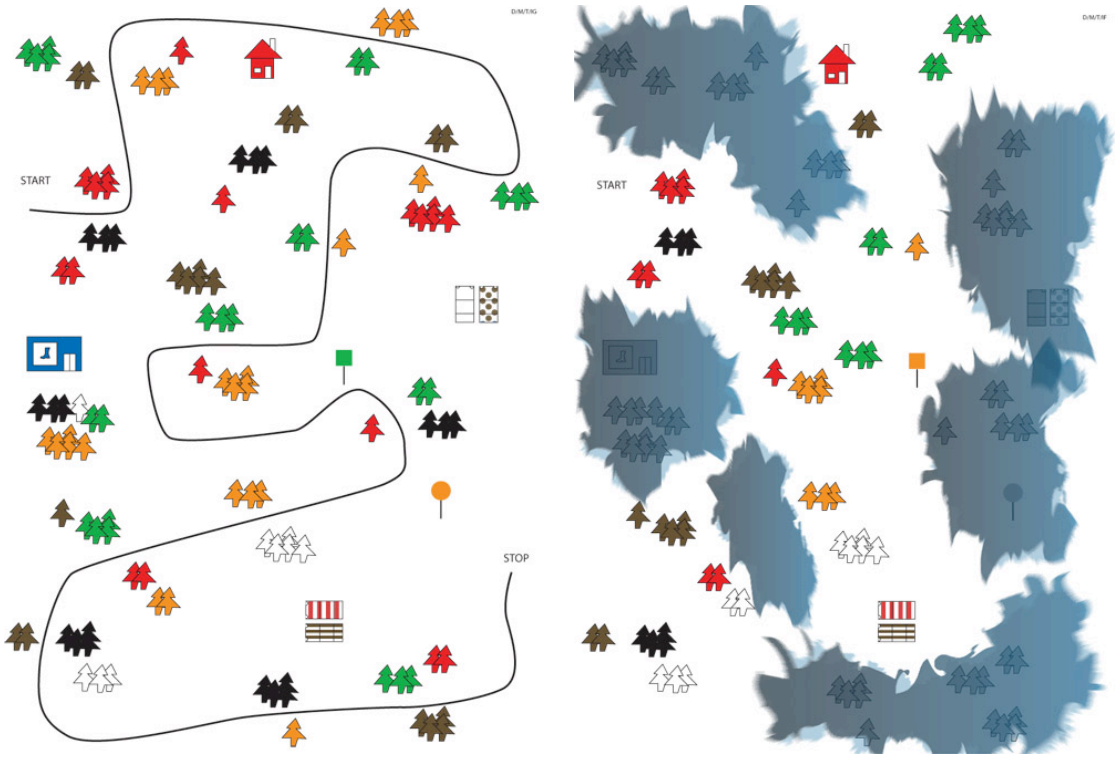


Figure 2: A pair of example maps; Instruction Giver left, Instruction Follower right

Marantz PMD670 recorder whereby Giver and Follower were recorded on two separate channels using two AKG C420 headset microphones. The speech was transcribed manually. The routes drawn by the Follower were recorded by the computer.

As the participants were in the same room, they could hear each other's speech. They could also see each other in the left half of their monitor, which showed the dialogue partner's upper torso video stream. The right half of the monitor showed the map.

Participants. In exchange for course credit, 64 undergraduates of the University of Memphis participated in pairs. In 4 dyads the participants knew each other previously.

3.3 Analysis and results

The recorded dialogues were coded for referring expressions. We present results for the first mentions of landmarks by the Instruction Giver. Introductory mentions should be both maximally independent of one another (as repeated mentions reflect precedence in naming a given object) and maximally detailed (as reductions in form characterise anaphora). Mentions of colour in landmark introductions were calculated as a proportion of opportunities

1. Over the course of single dialogues (by quartiles),
2. Across successive maps (1–8) and
3. Between those where the Instruction Giver lacked or already had experience as Instruction Follower.

The changes in the ratio of colour term use is depicted in Figure 3.

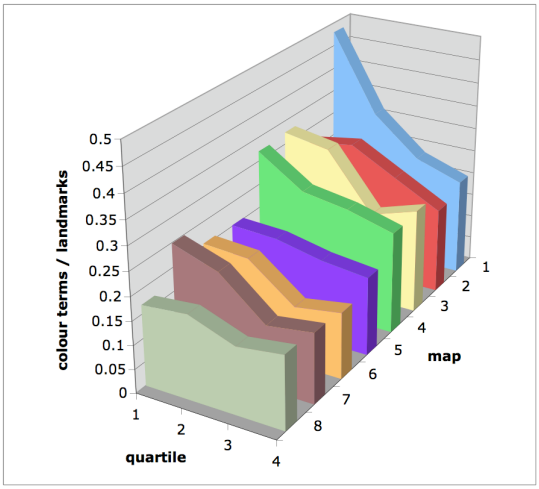


Figure 3: Change of the use of colour terms over quartiles of the eight maps

The use of colour terms significantly decreased over an average dialogue (effect of quartile within experience (2) x map encountered as Instruction Giver (4) x quartile (4) ANOVA on the arcsine transformed proportion of colour terms: $F_1(2, 54.8) = 15.57$, $p < 0.001$). Although there was no significant reduction across dialogues with the same Instruction Giver, the Givers used significantly fewer colour terms when they had served earlier as Follower (0.267 colour terms on average in the first four maps vs. 0.175 in the second four). This is a significant effect of experience ($F_1(1, 28) = 7.90$, $p < 0.01$).

Note that the orderliness of the ink blots on the Instruction Follower's maps did not have a significant effect. In contrast to colour, distinguishing features (number, kind, shape, pattern) are significantly more common in the maps where they are critical (used in more than 80%) and significantly increase within a dialogue. Thus, the decrease and low overall use of colour terms is not due to a general decrease in use of feature terms. There is also no effect of prior experience as Giver for useful features. The detailed results are presented in Guhe and Bard (2008).

3.4 Discussion

The participants adapted their use of colour to its low utility in the given task environment. The adaptation was distributed between speaker and listener. The use of colour terms does not fall significantly over the 4 dialogues a participant has the role of Instruction Giver, but there is a significant drop when the participants exchange roles: experience trying to match colour terms to grey-scale objects as Instruction Follower discourages to mention colour as Instruction Giver. Any listener-centric effect is outweighed or fuelled by a speaker-centric appreciation of utility.

4 Utility and task environment

4.1 Utility and selection probability

This is not the place to delve into the depths of the ACT-R theory, see Anderson (2007) for the most recent account. For the model described below it is only relevant that in ACT-R procedural knowledge (such as to decide whether to use colour or not) is encoded as production rules, or productions for short. A production is basically an if–then rule: *if* a certain set of conditions are given *then* execute a specified action.

In ACT-R, each production has a utility value. The utility is an estimate of how likely the use of the production results in achieving the current goal (here: successfully describing the landmark to the interlocutor).

Productions' utilities are important in the cases in which more than one production is applicable for a given set of conditions. Then, the utilities serve to compute the probabilities with which a production is selected. This selection probability is computed as:

$$P_i = \frac{e^{U_i / s}}{\sum_j e^{Uj / s}}$$

with:
P_i: selection probability for production i
U_i: utility of production i
s: noise in the utilities (defaults: $s = 1$)
j: set of all applicable productions (including i)

Utility values are learnt over time. After a production has been used, its utility is updated depending on whether it was successful according to the following equation:

$$U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)]$$

with:
U_i: utility of production i
n: number of applications of the production
α: learning rate
R: reward

If the production is applied successfully, the utility is updated with a positive reward, if it is unsuccessful, it receives a negative reward.

Anderson (2007, p 161) points out that this is basically the Rescorla-Wagner learning rule (Rescorla and Wagner 1972) or the delta rule by Widrow and Hoff (1960). So there is nothing special 'ACT-R-ish' about this rule; it is a general learning rule.

4.2 Structure of the task environment

In the maps about half of the landmarks on the Instruction Follower's map are obscured by ink blots, and, therefore, don't have colour. Additionally, some of the route critical landmarks mismatch in colour. Overall this means that using colour to describe a landmark is successful in only about 40% of cases. By comparison, using the distinguishing feature of a map is successful in about 92% of cases.

5 Model

5.1 Introduction

The following analyses compare the model's performance to the introduction of the first 33 landmarks of each map by the Instruction Giver. The 33rd landmark is still mentioned in 206 of the possible 256 cases (32 dyads with 8 maps each). The 34th landmark is introduced only 186 times.

There are three main patterns in the data. Firstly, map 1 behaves differently than the other maps in that the number of colour terms shows a pronounced drop from 0.6 to 0.25 (taken from the means of the first and last three values). Secondly, maps 2 to 4 each show a decrease of colour rate from 0.3 to 0.2. Thirdly, in maps 5 to 8 – after the role change – the colour rate drops in each map from 0.2 to 0.15. (This lower colour rate is the basis for the effect of role change.)

Thus, between maps the colour rate is going up again. Explanations may be that the longer-term utility of colour (learnt over a lifetime) or the textual prompting between dialogues exert some influence. The fact that the colour rate in maps 5 to 8 starts at the same rate as it ends in maps 2 to 4 may be due to the utility learning during the time as Instruction Follower. But a more detailed model is needed to explain this.

5.2 The model

The model is not a fully implemented ACT-R model, but just uses the two equations for updating production utility and probability of production selection introduced above. The model contains two competing 'productions' one for using colour, one for not using colour. Because the Instruction Giver always has colour available to describe a landmark, the model assumes that both productions are applicable for each landmark. Thus, the model is similar to the ACT-R model for an experiment by Friedman et al. (1964), described by Anderson (2007, p. 165–169; in this experiment participants have to predict which one of two lights will be lit.) Using the other features would be modelled as analogous sets of productions.

For each decision, the model selects one of the productions according to their utilities and corresponding selection probabilities at that time. After the decision has been made, the usefulness of colour is determined according to the structure of the task environment (thus, using colour is successful in 40% of cases) and the utility of the selected production is updated accordingly. For a successful application the production receives a reward of $R = 14$; if it is unsuccessful it receives a reward of $R = 0$ (cf. Anderson 2007, p. 162).

The results reported in the remainder of this section were obtained by 500 runs of the model. However, just 32 runs – matching the number of dyads in the experiment – suffice to get significant results; more runs of the model just produce a smoother curve.

5.3 Map 1

For the first map the model starts with the following estimated utilities:

$U_{\text{colour}}(1) = 5.5$
$U_{\text{no-colour}}(1) = 5$

These values mean that the colour-production has a probability of being selected of 0.622, which is close enough to the mean of the first three values of 0.594. (Using $U_{\text{colour}}(1) = 5.4$ would give an initial probability of 0.599, but one can be too fussy.)

The final average utilities are:

$U_{\text{colour}}(33) = 4.6$
$U_{\text{no-colour}}(33) = 7.7$

Choosing these initial utilities gives an excellent fit to the data, see Figure 4. A regression using the model as predictor for the data shows a significant correlation ($\beta_1 = 0.90$, $p < 0.001$) that accounts for 72% of the variance ($R^2 = 72\%$, $F(1, 31) = 79.5, p < 0.001$).

However, the initial values are not that important, and the model matches the data significantly for a wide range of start values, as long as $U_{\text{colour}}(1) > U_{\text{no-colour}}(1)$ and the values are not close to the extremes of 0 and 14. The same holds for the following simulations.

5.4 Maps 2 to 4

For maps 2 to 4 (see Figure 5) the initial utilities were set to:

$U_{\text{colour}}(1) = 5.5$
$U_{\text{no-colour}}(1) = 6.5$

resulting in final average utilities of:

$U_{\text{colour}}(33) = 4.5$
$U_{\text{no-colour}}(33) = 7.5$

The regression shows that the model accounts for 66% of the variance (R^2 = 66.3%, $F(1, 31) = 61.0$, $p < 0.001$) with $\beta_1 = 2.44$ ($p < 0.001$).

5.5 Maps 5 to 8

Finally, for maps 5 to 8 (see Figure 6) the initial utilities were set to:

$U_{\text{colour}}(1) = 3$
$U_{\text{no-colour}}(1) = 4$

resulting in the final average utilities

$U_{\text{colour}}(33) = 3.3$
$U_{\text{no-colour}}(33) = 7.7$

The model accounts for 52.7% of the variance (R^2 = 52.7%, $F(1, 31) = 34.6$, $p < 0.001$) with $\beta_1 = 0.84$ ($p < 0.001$).

6 Conclusions

There are two main conclusions from the research presented here. Firstly, the dialogue partners indeed adapt their naming behaviour to the task environment. More specifically, they adapt to the fact that colour is an unreliable distinguisher for the landmarks on the maps. (This is amplified by the fact that the participants do not make a substantial effort to identify the parts of the maps that are obscured by ink, which shows in the absence of an orderliness effect.)

Secondly, the simple computational cognitive model accounts for this change. In particular, the model shows that the change in behaviour is indeed an adaptation to the structure of the task environment, because the rate of the probabilities and the changes in the probabilities with which

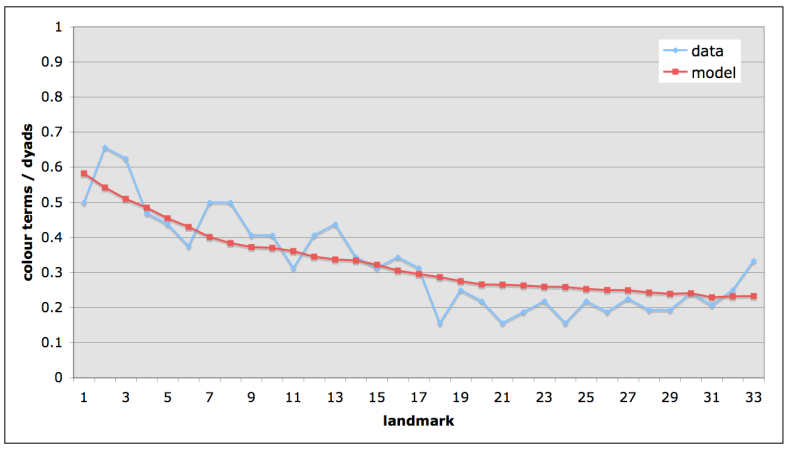


Figure 4: Comparison of data and model for the first 33 landmarks in map 1.

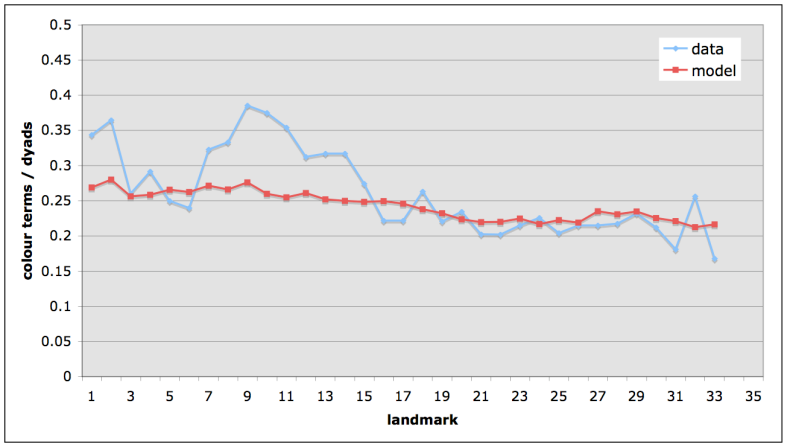


Figure 5: Data and model for maps 2 to 4.

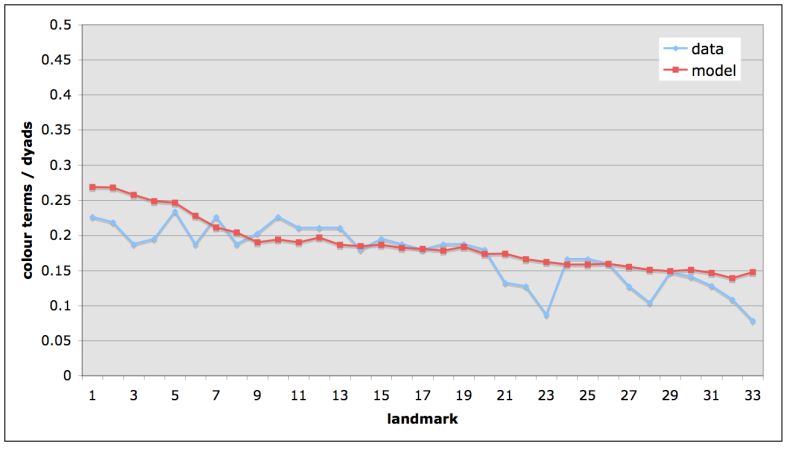


Figure 6: Data and model for maps 5 to 8.

colour is used as a descriptor is a direct result of the fact that colour can be successfully used for about 40% of the landmarks on the maps. Thus, rational analysis (the fact that memory reflects the probabilities encountered in the environment) explains the observed phenomenon.

Although – after the fact – it may not be too surprising that rational analysis explains the observed phenomenon, the result is more far-reaching, because the influences of the task environment on naming behaviour (the generation of referring expressions) has not yet been reported.

7 Future work

Our future research will address a number of direct follow-up issues. Firstly, the model will be extended to account for the changes in the mentions of the distinguishing features (number, pattern, kind, shape). Secondly, after a more detailed analysis of the data we will extend the model to account for individual adaptation patterns in the sense that the model can account for groups of dyads showing similar dialogue histories. For this, we will model the landmark introductions made by the Instruction Follower as well. This model serves as starting point for a comprehensive ACT-R model of how referring expressions (including repeated mentioned of landmarks) are generated in the given task.

Acknowledgements

This research was supported by grant NSF-IIS-0416128 to Max Louwerse, Art Graesser, Mark Steedman, and Ellen Gurman Bard. Thanks to Antje van Oosten and Jonathan Kilgour for help with the data coding and extraction and to Max Louwerse, Nick Benesh, Gwyneth Lewis and Megan Zirnstein for providing the transcriptions.

References

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

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R. and Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6), 396–408.

Arnold, J. E. and Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, 56(4), 521–536.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M. P., Doherty-Sneddon, G., and Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42(1), 1–22.

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

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

Dale, R. and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2), 233–263.

Friedman, M. P., Burke, C. J., Cole, M., Keller, L., Millward, R. B., and Estes, W. K. (1964). Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (ed.), *Human Brain Function* (2nd ed.) San Diego, CA: Academic Press.

Garrod, S. and Doherty, G. (1994). Conversation, coordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53, 181–215.

Guhe, M. and Bard, E. G. (2008). Adapting referring expressions to the task environment. In: *Proceedings of CogSci 2008*.

Jordan, P. W. and Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Kronmüller, E. and Barr, D. J. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *Journal of Memory and Language*, 56(3), 436–455.

Paraboni, I., van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2), 229–254.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 98–110.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

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

Cultural Differences in Computer-Mediated Communication

Susan Fussell
Carnegie Mellon University

Computer-mediated communication (CMC) technologies provide new opportunities for people to converse across space and time. Today, people connect with others from around the world by participating in chatrooms and discussion lists, by joining global game communities and virtual worlds, by authoring and reading blogs with an international scope, and by a variety of other means. In the work domain, firms are establishing global teams with members from a diverse set of nations who meet via an array of media including audio, video and text. Bridging nations via technology does not, however, guarantee that the cultures of the nations involved are similarly bridged. Mismatches in social conventions, work styles, power relationships and conversational norms can lead to misunderstandings that negatively affect the interaction. For the past few years, my students and I have been exploring the ways in which culture influences computer-mediated communication. Our goal is to develop a theoretical understanding of how culture influences computer-mediated communication and to inform the design of new tools to enhance cross-cultural communication.

In this talk, I will first provide an overview of the theoretical framework guiding our work. Then, Ill present three examples of cultural differences that we predict will influence CMCcommunication styles, individualism vs. collectivism, and peripheral awarenessand describe the laboratory, field and survey studies we have conducted to test these predictions. Ill then discuss several projects we are doing that aim to improve cross-cultural communication through training cultural sensitivity and by intervening in dialogues when problems arise. Ill conclude with some thoughts about how research on culture and CMC can be extended into new domains, such as support for communication in developing regions.

Can Aristotelian Enthymemes Decrease the Cognitive Load of a Dialogue System User?

Ellen Breitholtz
Dept. of Linguistics
University of Gothenburg
Box 200, SE-405 30 Göteborg
Sweden
ellen@ling.gu.se

Jessica Villing
Dept. of Linguistics
University of Gothenburg
Box 200, SE-405 30 Göteborg
Sweden
jessica@ling.gu.se

Abstract

In-vehicle dialogue systems are gaining an increased interest in the automotive industry. Dialogue systems allow the driver to use her voice, instead of her eyes and hands, to control devices in the car and thereby increase safety. Although speech is a natural way of communicating, the dialogue itself might increase the cognitive load of the driver. In this paper we suggest a rhetorical perspective of dialogue management, using Aristotelian *enthymemes* to provide a model for analysing Information Redundant Utterances and discuss the implications this may have for in-vehicle dialogue systems.

1 Introduction

One principle according to which dialogue is managed is Grice's maxim of quantity (Grice, 1975), *do not make your contribution more informative than required*. This has often been interpreted as "make your contribution as short as possible", resulting in all utterances that may be deduced from the context or co-text being considered *Information Redundant Utterances* (IRUs). Walker (1996) points out that IRUs are often not redundant at all (thus actually adhering to the maxim of quantity rather than violating it) but instead serves to help lower the listener's cognitive load.

Using IRUs might be a way of releasing the user of a dialogue system from some of the cognitive load of the interaction itself. This can be of great importance, especially in some environments. In-vehicle spoken dialogue systems are gaining increasing interest since they enable the driver to perform secondary tasks (i.e. tasks not related to driving the vehicle) without having to take her eyes off the road or her hands from the steering wheel. Dialogue systems, unlike command based systems, also enable the driver to speak in a natural way, without having to memorise commands. The possibility of speaking freely and not having to navigate through a fixed menu structure is however not enough. Driving is a safety critical task where the driver has to concentrate on the driving (primary task) rather than the dialogue system (secondary task). Therefore it is crucial to minimise the cognitive load of the driver caused by the dialogue itself. A difficult question in this context is how to decide when to add an IRUs and when not to. Some redundancy may help relieve the working memory of the user of a dialogue system or an agent in a human-human interaction, while too much information will only increase the cognitive load. In this paper we will discuss how a rhetorical perspective may be of use in this balancing act, and suggest that *enthymemes*, as presented in Aristotle's *Rhetoric* (Kennedy, 2007), may provide a model for analysing these utterances.

The outline of the paper is as follows: First, we will discuss the notion of IRU, as presented by Walker (1996). We then suggest an approach to understanding IRUs inspired by Aristotelian rhetoric, especially the concept of *enthymeme*. In section 5 some empirical examples of arguments collected from a corpus of car-navigation instructions are presented and discussed. In section 6 we discuss the

relation between the enthymeme and cognitive load. Finally, some conclusions are drawn and an attempt is made to formulate an agenda for further research and name some possible application areas.

2 Information Redundant Utterances

A significant feature of natural dialogue is economy. This has been noted by many scholars in the fields of pragmatics and discourse studies, and given rise to such well known and generally accepted theories as that of implicature (Grice, 1975). Walker (1996) mentions Grice's maxim of quantity as an example of a generally assumed *redundancy constraint*. Utterances that violate the redundancy constraint are referred to by Walker as IRUs. An utterance is considered an IRU if it expresses a proposition that the listener can *retrieve from memory* or *infer*. Walker argues that the redundancy constraint is based on four assumptions about dialogue:

1. Unlimited working-memory: everything an agent knows is always available for reasoning;
2. Logical omniscience: agents are capable of applying all inference rules, so any entailment will be added to the discourse model;
3. Fewest utterances: utterance production is the only process that should be minimised;
4. No autonomy: assertions and proposals by agent A are accepted by default by agent B.

According to Walker the principle of avoiding redundancy has often taken precedence in work on dialogue modelling and overshadowed other factors that affect communicative choice. Walker presents corpus data in which agents frequently violate the redundancy constraint, which indicates that the fewest utterance assumption is not correct - sometimes other aspects of communication are more important than economy.

Walker's analysis of corpus data leads her to formulate three main functions of IRUs:

- To provide evidence supporting beliefs about mutual understanding and acceptance.
- To manipulate the locus of attention of the discourse participants by making a proposition salient.
- To augment the evidence supporting beliefs that certain inferences are licenced.

Let us now take a look at one of Walker's examples of IRUs. An utterance is produced by A to B while walking to work (Walker, 1996):

(1) A: i) Let's walk along Walnut Street
ii) It's shorter.

It is known to A that B knows that Walnut Street is shorter, so by the redundancy constraint A should only have said i). Walker claims that ii) is considered an IRU based on the assumption of 'unlimited working memory', i.e. that all knowledge and information an agent has access to is equally available at all times. Walker hypothesises that the mentioning of the well-known fact that Walnut Street is shorter is a way for A to ease Bs cognitive load.

Let us take a look at another of Walker's examples. The following exchange is taken from an discussion about individual retirement accounts.

(2) A: i) Oh no, individual retirement accounts are available as long as you are not a participant in an existing pension.
B: ii) Oh I see. Well [...] I do work for a company that has a pension.
A: iii) Ahh. Then you're not eligible for [the tax year of] eighty one.

Walker's analysis of this example is that iii) is considered an IRU based on the assumption that agents are logically omniscient, since B would have to apply an inference rule to conclude iii). The function of A's stating iii) is, according to Walker, to "augment the evidence supporting beliefs that certain inferences are licenced".

3 A Rhetorical Approach to IRUs

Much work on language usage in general and dialogue systems in particular has taken rhetoric into account. Two recent examples are Miller (2003), who discusses how the notion of rhetorical ethos is central in creating an agent that is capable of passing the Turing test, and Andrews *et al.* (2006) who focus on how social cues and emotion can make dialogue systems behave more naturally. A fruitful way of

incorporating the logos-part of rhetoric in linguistic theory is as starting point for frameworks for structural analysis. Hobbs (1985), Asher and Lascarides (2003), Mann and Thompson (1986) *et al.* have presented theories for understanding textual structure (Mann and Thompson) and utterance relations (Lascarides, Asher, Hobbs). However, in much of the literature on rhetorical relations, little attention is paid, as far as we know, to the way supposedly information redundant utterances serve to add new information to the discourse situation by pointing to a specific argument.

We would like to suggest a way of looking at IRUs that elucidates Walker's ideas about the functions of IRUs, and offers an alternative to the four assumptions of the redundancy constraint. The three functions of IRUs in Walker's study have in common that they aim to lead the listener to a certain conclusion, either by supporting a belief the listener already has, or by directing, or even redirecting, the attention of the listener. In other words - IRUs are rhetorical. Examples (1) and (2) are both illustrations of this. The fact that (1ii) is considered redundant according to the redundancy constraint seems to reflect not only the unlimited working memory assumption, but also the assumption that agents are non-autonomous and by default accept assertions and proposals by other agents. The relative autonomy of B makes it possible for B not to accept A's proposition. By providing a reason for choosing Walnut Street, A performs a rhetorical act that potentially increases the likelihood that the suggestion will be accepted by B. Example (2) also indicates that A wants to make sure that B draws a specific conclusion. It seems likely that A, if she did not find it of some importance that B draws the conclusion iii), might not bother to make the inference explicit - B could still be expected to make the inference. However, for B to do that would not necessarily make her "logically omniscient" – the assumption Walker (1996) claims to be the reason for considering (2ii) an IRU – just capable of making *some* inferences.

Interestingly, many of Walker's examples of IRUs and their respective antecedents constitute structures similar to that of an Aristotelian *enthymeme*. An enthymeme can be described as a logic-like deductive argument. In the *Rhetoric* (Kennedy, 2007), Aristotle claims that learned, scientific argumentation differs from practical, hands-on argumentation concerning every day matters: when you speak to people that are not experts in the area you are dealing with, and who do not have much experience with logical reasoning, it is, according to Aristotle, inefficient to present long chains of logical arguments. In persuasion he therefore recommends shortening the arguments, which results in them not being strictly logical. However, Aristotle still emphasises the logos-based, deductive nature of the enthymeme, and calls it "a sort of syllogism" (Kennedy, 2007). The premises needed to make an argument a "real" syllogism, is added by the listener from her knowledge of culture, situation and co-text (what has been said earlier in the speech or conversation), according to a "pattern" known as the *topos* of the enthymeme. This pattern can be very general assumptions based on physical parameters such as space (the small can be contained in the big), or more specific assumptions such as prejudice about people belonging to a certain group. The mentioning of one carefully chosen premise directs the attention of the listener in the direction that the speaker wants, and makes the listener a bit more likely to accept the proposition presented in the conclusion. The enthymeme might of course serve to persuade or even mislead a listener, but the same mechanism can also make it easier for an agent A to accept an honest and constructive proposal made by another agent, which would be helpful when quick decisions need to be made, or when A has to focus on some demanding parallel activity.

Let us go back to the colleagues walking to work. Example (1) above could easily be analysed within a rhetorical framework. Mentioning (1ii) could be a way for A to point to the argument about the shortest route, perhaps because they are running late. There could be other reasons to walk along Walnut Street, perhaps that it is more quiet. A might know that B usually prefers a busy street, but that she does not particularly like to walk, which would make the short-argument more persuasive. If they were not in a hurry, and A wanted them to walk along Walnut Street because it is nicer to walk along a quiet street than a busy one, A would probably say 'Let's walk along Walnut Street. It's more quiet' thus validating her suggestion. But it

is also possible that A would want to walk along Walnut Street for some reason that she does not want B to know about - for example because someone cute always walks his dog there at that time. So, even though she knows that B knows it is the shortest way to work, A still mentions it to point out the getting-to-work-on-time argument. The enthymematic argument looks something like this:

It's shorter
We want to go get to work on time

$\therefore$ Let's walk along Walnut Street

The "hidden premise", i.e. the premise that B adds to the argument, would be something that makes sense in the context, having to do with for example time (as above) or effort (we don't want to walk longer than necessary).The additional premise is necessary in order to make the enthymeme fit with the relevant topos. This is also true in the case of (2), where two premises are expressed, but the expressed premises do not logically entail the conclusion.

Individual retirement accounts are available
as long as you are not a participant in
an existing pension
I do work for a company that has a
pension

$\therefore$ (Then) you're not eligible for
eighty one

A rhetorical perspective that uses enthymematic arguments as an explanation model for how information is given and withheld, would be based on a different set of assumptions about dialogue than those Walker formulates as the basis of the redundancy constraint. Thus we propose the following rhetorical principles

1. Limited working-memory: suggestions help agents to reach a certain decision

2. Logical capacity: agents are capable of applying some inference rules, some entailments will be added to the discourse model;

3. Utterance production: should be balanced so as to maximise persuasion

4. Autonomy: assertions and proposals by agent A are not accepted by default by agent B, and different agents may or may not share goals and intentions.

4 Enthymemes in Car Navigation Instructions

In a data collection carried out within the DICO project (http://www.dicoproject.org)[1], a driver is given navigation instructions by a passenger, in between the instructions the passenger interviews the driver about personal matters such as favourite food, number of siblings, favourite holiday resort, etc. The aim was to study human-human in-vehicle conversation with respect to how humans adapt the way they speak to the cognitive load of the other dialogue partner. The data includes examples of enthymematic arguments, of which some are also IRUs according to Walker's definition.

(3) A: i) Vi håller höger här på (*Let's keep to the right here*)
ii) Så vi kan...byta (fil) (*So we can...change (lanes)*)

Example (3) is uttered by a passenger (who provides the driving instructions) in a situation where both driver and passenger know that it is time to keep to the right in order to be able to change lanes. The passenger has stated a minute or so earlier that they should change to the right lane. Considering the information the driver has about the traffic situation and the previous instructions given by the passenger, (3ii) should not be necessary according to the redundancy constraint. (3) can also be seen as an enthymeme:

So we can...change (lanes)
In order to change lanes we have to keep to the right

$\therefore$ Let's keep to the right here

(3i), the proposition that they should keep right, is the conclusion of the enthymeme and the explicit premise (3ii) (they want to change lanes). The

[1] DICO is a project that aims to demonstrate how state-of-the-art spoken language technology can enable access to communication, entertainment and information services as well as to environment control in vehicles. A priority in the project is cognitive load management for safe in-vehicle dialogue.

non-explicit premise is something like 'if we are to change lanes we have to keep to the right', which is a fairly general assumption about spatial relations - the topos of the enthymeme.

(4) *A*: i) Rosengatan ja det måste vara nästa (*Rose Street yes it has to be the next*)
ii) för vi kommer inte så mycket längre (*cause we don't get much further*).

In example (4) it is clear to the driver that the street is ending. By supplying the premise (4ii) he points to an enthymematic argument based on a number of premises, most of which have been stated earlier (for example that Rosestreet crosses the street they are driving down), and one that has to be inferred (if you know that a street crosses the one you are driving down, and you haven't yet past it, and there is only one street left, this has to be the street you are looking for).

We have also looked at data recorded for the purpose of a master thesis about car navigation instructions (Caroline Bergman, work in progress). In this case the instructions are given over the telephone by a person with access to maps and driving instructions on the internet.

(5) *A*: i) Ta till vänster vid Redbergsplatsen
ii) står det här ja. (*turn left at Redbergsplatsen it says here.*)

Example (5) demonstrates the need to motivate for rhetorical purposes rather than to provide new information about one's reasons. The driver is well aware that the instructor is using a map and written driving instructions to be able to help the driver navigate. Still the instructor repeatedly validates her instructions by stating that the map or other instructions 'says so'. It seems probable that the driver has reason to be suspicious of the instructions, since the person giving them is somewhere else and does not have access to any information about the traffic situation that the driver does not provide.

5 Cognitive Load and Efficiency

As humans we need reasons to validate propositions we are presented with. We know this intuitively – it is difficult to complete a task if we are just presented with single pieces of information that do not seem to be connected. The same conclusion can be made based on different premises, and we often want to know which argument the speaker is referring to before we accept a proposition. There are situations where the standard way to instruct is by single utterances (or orders), such as in the military, or in other contexts where the roles are very well defined, and the *modus operandi* of the activity well rehearsed, such as in surgery. We agree with Walker's conclusion that IRUs serve to ease cognitive load in different ways. Our hypothesis is that the reason why they do this is often because the enthymematic structure helps the recipient of the IRU to make up her mind - if the provided premise fits into an argument she finds acceptable she will agree with the proposition, if not she will disagree. Neys and Schaeken (2007) show that the tendency to make logical rather than pragmatic inferences increases when under heavy cognitive load, which indicates that pragmatic inferences use more working memory. This supports the idea that it would be good to present arguments in a form resembling a logical argument rather than just presenting the proposition - even if the recipient is aware of the information in the premise (IRU) provided.

6 Concluding Remarks

Studying in-vehicle conversations reveals that interacting with someone while driving is always distracting (see e.g. Patten *et al.* (2003)), and sometimes dangerously so. Conversation increases the cognitive load of the driver and thus prevents her from fully focusing on the primary task of driving. Studies of cell phone conversations have revealed that the major reason why cell phone conversation is dangerous is not the handling of the cell phone (i.e. the use of hands free cell phone is not safer than a manual cell phone), but the conversation itself increases the cognitive load of the driver to such an extent that the risk of an accident increases (Redelmeier and Tibshirani, 1997). Most user studies carried out to measure cell phone conversations impact on driving behaviour are carried out in car simulators, and the parallel task is to perform mental processing tasks such as arithmetic operations.

These studies point at a significantly decreased driving performance. Esbjörnsson *et al.* (2006) on the other hand, studied real cell phone conversations in cars driving in real traffic. They found that in human-human conversation the dialogue partners have strategies for dealing with distraction and increased cognitive load. Humans tend to 1) sense when a particularly stressful situation is coming up and adjust by, for example, pausing the conversation and 2) generally use the conversational "rules" that keep verbal interaction running smoothly – after all, keeping a conversation going in any situation adds to the cognitive load of the speakers. However, in the context of in-vehicle conversation with a dialogue system we can normally not expect this kind of adjustment. The system can be compared to a *remote caller* (Schneider and Kiesler, 2005), i.e. a dialogue partner not sitting in the passenger seat but speaking to the driver over the phone and thus does not have access to the traffic situation. Problem 1), that of detecting and managing particularly stressful situations, is not addressed in this paper. Instead we have focused on a way to potentially minimise the cognitive load that is caused by the conversation itself.

A rhetorical perspective provides a model for interaction that works for interactions in a context where the agents do not necessarily have a common goal or intention. The mechanisms that enables persuasion, can also be used in order to explain something in an easily comprehensible way. A skilled rhetorician is often also a skilled teacher, since it is easier to understand something if one understands the argument behind it. In the context of a dialogue system that is advanced enough to be able to handle conversation that is to a certain degree “free”, a rhetorical perspective would be beneficial. This would be the case for contexts when the system has an agenda distinct from that of the user, e. g. to make the user buy something or convince the user about the importance of a healthy lifestyle. In a domain such as car navigation, where user and dialogue system have a common goal, it might still be beneficial for the system to be able to provide a premise that points to an argument that would explain its reasons for giving a certain answer or instruction. Such premises would be helpful not only in situations where the system adds new information, such as if the user has asked for the quickest route and the dialogue system proposes a route that does not seem to be the shortest possible, and the system explains that some roadwork is going on or there is a one-way street along the shortest route, but also in situations where the contribution serves a rhetorical purpose rather than an informational one. The system’s pointing to an enthymeme relevant to the situation may make it easier for the driver to decide whether to accept the instruction or not. This potentially minimises the cognitive load since the driver has to make fewer inferences, but still is not overloaded with all the evident premises of strictly logical reasoning.

7 Future work

In future work, we plan to further analyse data collections carried out in the DICO project, and investigate how enthymemes and IRUs are used in human-human dialogues. In addition to this we would like to perform an experiment where subjects are instructed to solve an ethical problem online. Based on the results a repetition of the experiment could be performed where subjects are divided into two groups. In this second part of the experiment the conversations will be manipulated. One group will be provided premises such as were given by subjects in part one, the second group will get premises that do not make sense. This kind of experiment would allow us to compare the decision making capacity in the two cases. Hopefully it would also give information about when it is beneficial to motivate a proposition and what kind of information should be supplied. The results of DICO data analysis and experiments will possibly show some regularities similar to the notion of topoi, and might give an idea about which enthymemes a car navigation system should be able to point to.

Acknowledgements

The authors would like to thank Caroline Bergman and DICO (Vinnova project P28536-1) for providing data. DICO also in part supported Villing’s work.

References

P Andrews, S Manandhar, and M De Boni. 2006. Integrating emotions in persuasive dialogue: A multi-layer

reasoning framework. In *Paper presented at the 19th International FLAIRS , May 11-13*.

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

Mattias Esbjörnsson, Oscar Juhlin, and Alexandra Weilenmann. 2006. Drivers using mobile phones in traffic: An ethnographic study of interactional adaptation. *International Journal of Human Computer Interaction, Special issue on: In-Use, In-Situ: Extending Field Research Methods.*

H P Grice, 1975. *Logic and Conversation*, chapter Syntax and Semantics: Speech Acts, pages 41–58. Acad Press:NY.

Jerry R Hobbs. 1985. On the conherence and structure of discourse. Technical report, Center for the Study of Language and Information, Stanford University.

George A Kennedy. 2007. *On Rhetoric, a theory of civic discourse*. Oxford university press.

W C Mann and S A Thompson. 1986. Rhetorical structure theory: Description and construction of text structures. In *Proceedings of the NATO Advanced Research Workshop on Natural Language Generation, Nijmegen, The Netherlands, August 19-23, 1986*, pages 85–96. Springer.

Carolyn R Miller. 2003. Writing in a culture of simulation: Ethos online. In Martin Nystrand and John Duffy, editors, *Towards a Rhetoric of Everyday Life: New Directions in Research on Writing*, chapter Writing in a Culture of Simulation: Ethos Online, pages 58–83. University of Wisconsin Press.

Wim De Neys and Walter Schaeken. 2007. When people are more logical under cognitive load: dual task impact on scalar implicature. *Experimental psychology*, 54(2):128–133.

Christopher J. D. Patten, Albert Kircher, Joakim Östlund, and Lena Nilsson. 2003. Using mobile telephones: cognitive workload and attention resource allocation. *Accident Analysis & Prevention*, 36:341–350.

D A Redelmeier and R J Tibshirani. 1997. Association between cellular telephone calls and motor vehicle collisions. *Nwe England Journal of Medicine*, 336:453–458.

Mike Schneider and Sara Kiesler. 2005. Calling while driving: effects of providing remote traffic context. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 561–569, New York, NY, USA. ACM Press.

Marilyn A. Walker. 1996. The effect of resource limits and task complexity on collaborative planning in dialogue. *Artificial Intelligence*, 85(1-2):181–243.

Leveraging Minimal User Input to Improve Targeted Extraction of Action Items

Matthew Frampton, Raquel Fernández, Patrick Ehlen, Anish Adukuzhiyil and **Stanley Peters**
Center for the Study of Language and Information
Stanford University
{frampton,raquel,ehlen,ajohna,peters}@stanford.edu

Abstract

In face-to-face meetings, assigning and agreeing to carry out future actions is a frequent subject of conversation. Work thus far on identifying these action item discussions has focused on extracting them from entire transcripts of meetings. Here we investigate a *human-initiative targeting* approach by simulating a scenario where meeting participants provide low-load input (pressing a button during the dialogue) to indicate that an action item is being discussed. We compare the performance of categorical and sequential machine learning methods and their robustness when the point of user input varies. We also consider automatic summarization of action items in cases where individual utterances contain more than one type of relevant information.

1 Introduction

Regrettably, people do not always pay attention to everything you say. In fact, research on lexical change blindness suggests they miss more than you might imagine (Sanford et al., 2006). But such attention-constraining strategies can prove adaptive in the face of so-called "information overload," and the myriad pressures on attention that arise from living in the modern era (or, perhaps, any era). For example, an effective attention strategy during a business meeting might be to pay close attention to e-mail on your laptop while processing the ongoing meeting dialogue in a shallow way that picks up on segments of interest to you, or in which it seems you are about to be assigned some task. When those patterns of dialogue arise, you then pay closer attention to the dialogue, or even participate yourself. Such strategies come second nature to us.

But this strategy of targeted listening can be employed in machine interpretation of meeting dialogue as well, using an approach to dialogue processing we call *targeted understanding*. While a machine's interpretation of semantics in multi-human dialogue faces different obstacles from those faced by a human—lacking the facility with context and intentionality that we take for granted—the general approach to interpretation can be similar: Only segments that contain certain patterns of dialogue are identified as deserving close attention, followed by a deeper semantic analysis of those segments for the most relevant bits of information.

In this paper we briefly discuss how we use targeted understanding to identify the tasks people agree to in meetings (their *action items*) from multi-party meeting dialogue. Work thus far on this endeavor has focused on extracting action items from entire records of meetings (Purver et al., 2007; Ehlen et al., 2008), relying on a machine-initiative approach that extracts all possible action item discussions and then asks meeting participants to cull them after the meeting is finished. Here we will steer a slightly different tack, investigating the potential of "human-initiative targeting" that allows participants in a meeting to give some indication of an area of interest—by, say, pressing a button when an action item is being discussed. We then use automatic methods to extract the semantic properties of utterances that are salient to that segment of dialogue, and generate a readable summary.

In the next Section we describe previous work on extracting action items. After that, in Section 3,

we present our approach and methodology for this study. Sections 4 and 5 present our experiments and results: First with respect to the task of detecting those utterances that contain semantic information related to action items; and second with respect to extracting different kinds of properties from single utterances that contain more than one type of action item-related semantic information. We conclude with directions for future work in Section 6.

2 Targeted Understanding of Action Items

The process of assigning and agreeing to carry out future actions frequently arises through some channel of communication, such as e-mails or dialogue. They are often called *action items* or *next actions* and arise as *public commitments* to undertake a task. Several recent efforts have sought to utilize this communicative channel to extract them automatically, and to mine and summarize useful information from them.

2.1 Action Items in Dialogue

How do people in meetings discuss action items? Because the process of deciding what tasks will be done and who will do them is a common and significant interaction during meetings, their discussion approximates an *exemplary structure*, adhering to a recognizable pattern—even if that pattern comes spread over several persons and multiple utterances.

(1) A: We should have a rerun of the three of us sitting together
B: Sure
A: Some time this week again
C: OK
A: And finish up the values of this
B: Yeah

In the first place, there is usually some discussion of the task that needs to be performed. In the example above the sub-utterances *"have a rerun of the three of us sitting together"* and *"finish up the values"* contribute to a *task description*. The first utterance also includes a second component that is commonly found in action items, which is discussion of who will be responsible for—or take *ownership* of—the task to be performed (in this case, all participants, or *"we"*). A third component is some designation of the *timeframe* in which the task should be completed, in this case *"some time this week"*. Finally, because this is a public, joint commitment and not a solitary one, one often hears some indication of agreement from the participants agreeing to the commitment (*"Sure", "OK", "Yeah"*) Because acknowledgments like these help to glue together verbal acts of coordination, *agreement* is an important fourth component in such discussions.

Thus, a dialogue that discusses an action item tends toward some approximation of this exemplary structure, and includes utterances that play one or more of these four roles at a time. Granted, the structure is exemplary, so sometimes one of these elements (such as the timeframe) may not be present. But in general, the closer a round of dialogue comes to representing these four types of dialogue moves—*task description*, *ownership*, *timeframe*, and *agreement*—the more likely we find that some future task or action item is being discussed.

2.2 Structural Extraction Approach

This structural insight was fleshed out in Purver et al. (2006; 2007). Others (Morgan et al., 2006; Hsueh and Moore, 2007) had attempted a *flat* approach to action item detection in dialogue where utterances were simply marked as either being relevant to an action item discussion or not. Purver et al. (2007) replaced this flat classification approach with a structured, hierarchical one. They trained four linear Support Vector Machine (SVM) classifiers to detect utterances that correspond to each of the four Action Item-related Dialogue Acts (AIDAs) in Table 1. Then they used a *super-classifier* trained with the hypothesized labels and confidence scores of the four independent classifiers to detect clusters of those sub-classes, which indicate probable discussions of action items. On the task of detecting action item discussions, this approach achieved an F-score of 0.45, (using a criterion of at least 50% overlap between hypothesized and oracle action item discussion), compared to 0.35 using a flat approach with the same feature and data sets.

The strategy of attending to and targeting a specific dialogic structure exhibits a clear benefit over a flat approach. But note that this approach to hierarchical classification does not presume any sequential dependencies in the utterances, since they are classified separately and aggregated by window, thus

D	*description*	discussion of the task to be performed
T	*timeframe*	discussion of the required timeframe
O	*owner*	assignment of responsibility (to self or other)
A	*agreement*	explicit agreement or commitment

Table 1: Action item dialogue act (AIDA) classes.

ignoring any temporal organization that might exist in the exemplary pattern of action item discussions. This is one possibility we intend to investigate here.

2.3 Exploiting User Feedback

Another way to improve detection of action item discussions and their associated AIDAs is to involve a person in the loop who can provide some feedback about whether or not the detected utterances really do correspond to discussion of an action item.

This possibility was explored by Ehlen et al. (2007; 2008), who used a post-meeting browser tool to present detected action items to meeting participants taken from the DARPA CALO 2007 CLP evaluation. After each meeting, participants could review their action items, changing the *task description* (D), *timeframe* (T), and *owner* (O) entries in ways that allowed feedback to three of the four corresponding AIDA sub-classifiers. When users added action items to their to-do lists or rejected them, feedback for the super-classifier was also harvested.

These data from human feedback were used to re-train each of the targeted classifiers, allowing an assessment of whether implicit user feedback could help improve the models. Indeed, this type of feedback yielded F-score error reductions between 20 and 40% for different meeting sequences, indicating that human feedback could be useful.

Results such as these bring up the question of whether some other types of human input might yield similar improvements. Instead of requiring meeting participants to review action items after a meeting is finished, perhaps they could "mark" relevant segments of a meeting as they happen, by, for example, pushing a button when something occurs that corresponds to information they wish to recall or have extracted. Our first experiment in Section 4 simulates just such a scenario.

2.4 Summarizing Action Items

There is a growing interest in dialogue summarization, with most approaches attempting to summarize the content of entire dialogues (Zechner, 2002; Murray et al., 2005; Murray and Renals, 2007). The most obvious application of identifying action item discussions and their corresponding dialogue acts is to produce a more structured and targeted meeting summary by providing a descriptive record of the tasks assigned, perhaps presented as an automatically generated to-do list.

Purver et al. (2007) made a preliminary attempt at generating extractive summaries of action items, focusing on utterances tagged as performing one of two AIDAs: either the *task description* (D) or the *timeframe* (T) during which the task is to be performed. Their approach involved parsing the word confusion network (WCN) for each relevant utterance using a general rule-based parser (Dowding et al., 1993), which produced multiple short fragments rather than one full utterance parse. An SVM classifier was then trained to learn a model which ranked these phrases according to their likelihood of appearing in a gold-standard extractive summary. Various features were used including WCN, parse, lexical and temporal expression tags.

This approach produced mixed results. While precision was higher than that of a baseline that used the entire 1-best utterance transcription, only the F-scores obtained for *timeframe* outperformed the baseline. Besides yielding mixed results, this prior work did not consider summarisation of action items where utterances are tagged with multiple AIDA classes. In such cases, it is necessary to determine which bits of information are related to which dialogue act, and as a result the summarization task becomes more complicated. Our second experiment in Section 5 addresses this issue.

3 Approach & Methodology

The work of Purver et al. (2007) has shown that automatically identifying AIDAs in transcripts of full meetings is a difficult task—achieving F-scores below 0.25 (see Table 2). One reason is that AIDAs are

very sparse, making up only around 1.4% of utterances in a meeting transcript. In the first of two experiments, we want to investigate how on-line input given by meeting participants can reduce the sparseness problem and thus help in automatic identification. If participants could indicate where an action item is being discussed by, for instance, pressing a button during the ongoing dialogue, such "human-initiative targeting" could help the system to bypass large sections of dialogue in favor of specific, relevant regions.

We simulate participants' input by selecting sections of dialogue that include discussion of action items, and then use machine learning on the targeted sections to identify the AIDAs. In doing so we address a number of issues.

First, we investigate the degree to which human-initiative targeting can improve classifier performance by training only on windows of utterances instead of full meetings. The average length of an action item discussion is 7.8 utterances, and 92% of action items are at most 15 utterances long. Hence we allow the system to have access only to 15 utterance windows.

Secondly, we compare the performance of a Support Vector Machine (SVM) categorical classifier, as used by Purver et al. (2007), against a Hidden Markov model (HMM). The HMM is a sequential model, and so assuming that action item discussions exhibit regularities in sequences of utterance types, it may perform better.

Thirdly, we also investigate how robust classifier performance is with regard to when the human input is given. We first consider a case in which participants always press a button right at the end of an action item discussion, and then look at a presumably more realistic case in which participants may press the button at different times in relation to the end of an action item discussion. This allows us to investigate the extent to which performance degrades in less systematic and more realistic situations.

Our second experiment is concerned with extracting words from AIDAs that can be used to generate a useful descriptive summary of an action item discussion. As mentioned in the previous section, the fact that utterances can be tagged with multiple AIDAs complicates the task of extracting information for summarization purposes, since we need to distinguish between bits of information related to different AIDAs but contained within a single utterance. We address this issue in Section 5, focusing on those utterances that have been simultaneously tagged with classes D and O. Again, we compare performance of categorical (SVM) and sequential (HMM) classifiers.

For our two experiments, we used the ICSI Meeting Corpus (Janin et al., 2004), which contains recordings and manual transcriptions of naturally occurring research group meetings. In particular, we used the annotated sub-corpus of Purver et al. (2007), which consists of 18 ICSI meeting transcripts annotated using the AIDA classes shown in Table 1. The annotations also include a summary description for every instance of an AIDA class, created by manual selection of words and phrases from the gold-standard transcripts.

4 Experiment I: Targeted AIDA Detection

In this section we present our experiment on detecting AIDAs from targeted regions of meeting transcripts.

4.1 Data

The 18 ICSI meetings in our subcorpus have been annotated with 190 action item discussions in total (10.6 action items per meeting on average). To simulate user input, we generated two different data-sets from this corpus: a *systematic input* data-set and a *non-systematic input* data-set. The systematic input data-set was generated by extracting 190 sections of 15 utterances, and for each the last utterance corresponded to the last AIDA of an action item. This data-set simulates a scenario where participants always press a button right at the end of an action item discussion. The non-systematic data-set simulates a more realistic situation where user input is given at random points towards the end of an action item discussion. Here we allow the system to look 10 utterances backwards and 5 forward from the point when the input is given. The data-set was generated by extracting 190 sections of 15 utterances, where the input is assumed to be randomly given either immediately after the last AIDA of an action item discussion, or 1, 2, 3 or 4 utterances earlier.

Targeting sections of dialogue that contain action

item discussions obviously reduces AIDA sparseness considerably. Averaging over the *systematic* and *non-systematic input* data-sets (which are very similar in this respect), 13.7% of utterances (around 2 on average per window) are tagged with class D, 4.4% (around 0.6 per window) are tagged with class T, 9.5% (around 1.4 per window) are tagged with class O, and 14.6% (around 2.2 per window) are tagged with class A.

4.2 Classifiers & Features

We use the linear-kernel support vector machine classifier SVM*light* (Joachims, 1999) and the structural support vector machine classifier SVM*hmm* (Altun et al., 2003), which trains models that are isomorphic to hidden Markov models.

We train four individual SVM classifiers—one for each AIDA class—and compare their performance to that of one single HMM classifier that uses six different labels for the model states: labels D, T, O, and A for each of the AIDA classes, plus a label X for utterances outside the action item discussion and an insertion-class label I for those utterances inside an action item discussion that do not belong to any AIDA class. In all cases, we evaluate performance using 18-fold cross-validation, with each fold containing those 15-utterance windows that belong to the same meeting.

To train the classifiers, we use similar features to those of Purver et al. (2007), derived from the properties of the utterances in context: lexical unigrams, durational features from the transcriptions, dialogue act tags from the ICSI-MRDA annotations (Shriberg et al., 2004), temporal expression tags using the MITRE TIMEX tool, as well as contextual features consisting of the same features for the immediately preceding and following 5 utterances.

4.3 Results

The results reported in Purver et al. (2007) for the task of identifying AIDAs from whole meetings are shown in Table 2. Using simulated participant input to target regions of dialogue that contain action item discussions, we are able to improve these baseline results by more than 30% (see Table 3).

Table 3 shows the scores we obtained when simulated participant input was provided, systematically at the end of an action item discussion and non-systematically at any point in the second half of the discussion. In this case the results for these two different data-sets are very similar. The non-systematic input data-set yields slightly lower F-scores, but the drop is only statistically significant for classes D and A ($p < 0.05$ on a paired t-test). The slightly lower results may be due to the fact that some AIDAs may not fall into the 15 utterance window the classifier is looking at (for instance, if the input is given at the end of the action item and the discussion is more than 10 utterance long, then since the classifier is only looking 10 utterances back, the AIDA(s) at the beginning of the action item discussion are not considered), which reduces the number of available positive examples.[1]

	D	T	O	A
Recall	.19	.15	.21	.18
Precision	.18	.46	.27	.16
F-score	.19	.22	.24	.17

Table 2: SVMs trained on whole meeting transcripts

	D	T	O	A
Recall	.66	.57	.66	.78
Precision	.51	.45	.51	.49
F-score	.57	.51	.57	.60
Recall	.56	.52	.62	.82
Precision	.45	.45	.50	.44
F-score	.50	.48	.55	.57

Table 3: SVMs trained on targeted regions; systematic input (top) vs. non-systematic input (bottom)

Table 4 shows the results we obtained when we used a single HMM instead of four independent SVM classifiers. While recall is significantly lower for all classes ($p < 0.05$) leading to a drop of F-scores, the sequential model is able to achieve good precision results. In contrast to the SVMs, however, using the non-systematic input data-set with the HMM classifier leads to a statistically significant drop in performance, especially for classes A and D, where both recall and precision decrease ($p < 0.01$). This is perhaps not surprising, since the variability of the non-systematic data-set disrupts the sequen-

[1] A possible way of compensating for this would be to increase the size of the window. This however is not an optimal solution since the bigger the window the sparser the AIDAs.

	D	T	O	A
Recall	.48	.33	.45	.54
Precision	.53	.52	.50	.53
F-score	.50	.40	.48	.53
Recall	.32	.22	.32	.38
Precision	.45	.41	.46	.40
F-score	.37	.29	.38	.39

Table 4: HMM trained on targeted regions; systematic input (top) vs. non-systematic input (bottom)

tial organization that drives this kind of model.

While lexical features were the most useful in all cases, we observed that using the MRDA dialogue act tags `commitment` and `suggestion` improved precision significantly, especially for classes O and D. TIMEX tags boost scores for class T, although using targeted regions does not improve precision for this class.

In summary, using online input to target regions of dialogue where an action item is being discussed can improve AIDA detection substantially when compared to a no-input approach, even if the input is given randomly towards the end of the action item discussion. Although the sequential model yielded good precision scores, its performance was less robust to non-systematic user input. A possible reason for its lower recall even with the systematic data-set is that HMMs may not be so well suited when target classes are sparse:[2] if the model fails to hypothesize one AIDA where it should, it may then fail to hypothesize subsequent AIDAs. SVMs do not have this problem because each utterance is assessed independently.

5 Experiment II: Summarization of Utterances Tagged with Multiple AIDAs

Having indentified the constituent utterances in an action item, the next task is to summarize their action item-related semantic content so that it can be presented in a to-do list for the user. Here, we use a different methodology from Purver et al. (2007) that does not require a parser, and concentrate on extracting summary-worthy words from utterances that have been tagged with multiple AIDAs. While in general there is a large degree of independence between class distributions (with most cosine distances below 0.3), classes D and O often overlap, yielding a between-class cosine distance of 0.55 (where 1 represents exact correlation and 0 total independence). Hence we concentrate on those utterances that have been tagged as both *ownership* (O) and *task description* (D).

5.1 Methodology

In our 18 meeting corpus there are 162 utterances that have been tagged as both D and O. These utterances contain a total of 2697 words, 409 of which have been annotated as summary-worthy for class O, and 1015 as summary-worthy for class D. Example (2) shows a D + O utterance with the gold-standard summary-worthy phrases indicated in square brackets.

(2) It would be great if [*you*]$_O$ could um not transcribe it all but uh [*pick out some stuff*]$_D$

We use gold-standard extractive summaries as targets and train a classifier to decide whether or not each word in the manual transcription of a D + O utterance is summary-worthy for classes O and D, respectively. This approach exploits the fact that critical phrases that contain summary-worthy information for different AIDAs display characteristic syntactic, semantic, and lexical features.

To train our classifiers we used lexical trigrams (including the current word, and the immediately preceding and following words) and Part-of-Speech (PoS) tags generated by the Stanford PoS tagger (Toutanova and Manning, 2000). In all cases, testing was performed using 10-fold cross-validation. We experimented with the following types of classifiers:

- SVM: Two independent classifiers each trained to distinguish O and D words, respectively, from other words.
- SVM (O/D): One classifier trained to distinguish between O, D, and other words.
- HMM (B/I): One classifier trained to distinguish between O words (beginning and inside of sequence), D words (beginning and inside of sequence), and other words.[3]

[2] As mentioned in Section 4.1, AIDA classes in targeted regions make up between 4.4% and 14/6% of utterances.

[3] The end of the sequence is labelled with the inside (I) tag.

5.2 Evaluation

We evaluated each classifier's performance against the manually-annotated summary descriptions. Recall was therefore the proportion of words in the gold-standard summaries which overlapped with the words extracted by the classifiers; precision was the proportion of words extracted by the classifiers which also appear in the gold-standard summaries.

The O and D classes are compared to different baselines. Since the role of the O class is to assign responsibilty for a task, a large number of utterances tagged with O contain names or pronouns identifying the responsible party. Hence it is reasonable to use a baseline which tags all instances of first and second person personal pronouns (*I, you, we*) as positive. For class D, there was no clear majority POS class, so we settled on a baseline that tagged half of all words in D utterances as positive, where this half was selected randomly.

5.3 Results

Table 5 shows results for the different classifiers. All of the classifiers achieved substantially higher F-scores than the baseline for both *ownership* (O) and *task description* (D).

	Ownership			Description		
Model	Re	Pr	F1	Re	Pr	F1
Baseline	.39	.59	.47	.53	.38	.44
SVM	.76	.56	.64	.80	.64	.71
SVM (O/D)	.61	.67	.64	.74	.68	.71
HMM (B/I)	.61	.69	.65	.74	.71	.73

Table 5: Extraction of summary-worthy O/D words

For O, all of the classifiers achieved very similar F-scores. However a t-test shows that the HMM's score is significantly higher than the SVM(O/D) ($p < 0.005$). For D, the HMM performed significantly better than both the SVM(O/D) and SVM(D) classifiers in terms of precision and F-score ($p < 0.01$). Its F-score of .73 is much higher than that achieved by the best model of Purver et al. (2007): .38, lower even than their baseline which was the entire 1-best utterance transcription (see Section 2.4). Although those results are not directly comparable to ours, (since we used gold-standard transcriptions rather than WCNs, and focused on utterances that had been tagged with 2 rather than 1 AIDA class), we believe they show that the general approach has promise, and that the sequential model is well-suited to this task.

6 Conclusions & Future work

We have simulated a "human-initiative targeting" approach to action item detection where participants provide input—e.g. by pressing a button—to indicate that an action item is being discussed, which allows a system to concentrate on relevant dialogue regions. As a result we were able to improve the detection of action item-related dialogue acts (AIDAs) very substantially, obtaining F-scores that are twice as high as when using whole meetings.

Categorical models (SVM) proved to be more useful than sequential ones (HMM) for this task. The HMM yielded good precision scores but significantly lower recall, and so the overall performance was lower for this type of classifier. When we compared systematic user input given at the end of an action item discussion with less systematic input given randomly at different points towards the end of the action item, we found that the SVMs were more robust than the sequential model. This is not surprising since such unsystematic behavior disrupts the sequential organization which the HMM relies on.

We also addressed the task of extracting summary-worthy information from utterances that had been tagged with two AIDAs—*ownership* and *task description*—and found sequential models to be useful for this task, achieving F-scores of .65 and .73, respectively.

In the future we plan to experiment with a two-stage classification approach. This would involve first using SVMs to make classifications and provide confidence scores independent of sequence, and then second, giving this information to a sequential model that makes the final classifcations. Combining the two different types of classifier in this way may produce better results for both AIDA classification and summarization.

Our findings with respect to targeted understanding are useful, but of course, real user behavior during actual meetings will differ in many respects, and will surely prove more variable than what we have simulated here. Bearing this in mind, fu-

ture work will involve conducting an experiment in which we ask actual meeting participants to provide live button-pushing input during meetings when it occurs to them that an action item is being discussed. Only then can we know whether the approach described in this paper will be robust enough to handle the vagaries of real human behavior.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-07-D-01850004. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA, or the Air Force Research Laboratory.

References

Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Hidden Markov support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*.

John Dowding, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran. 1993. GEMINI: a natural language system for spoken-language understanding. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Patrick Ehlen, Matthew Purver, and John Niekrasz. 2007. A meeting browser that learns. In *Proceedings of the AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants*.

Patrick Ehlen, Matthew Purver, John Niekrasz, Kari Lee, and Stanley Peters. 2008. Meeting adjourned: Offline learning interfaces for automatic meeting understanding. In *Proceedings of the International Conference of Intelligent User Interfaces*, Canary Islands, Spain.

Pey-Yun Hsueh and Johanna Moore. 2007. Automatic decision detection in meeting speech. In *Proceedings of MLMI 2007*, Lecture Notes in Computer Science. Springer-Verlag.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Marcías-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. 2004. The ICSI meeting project: Resources and research. In *Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop*.

Thorsten 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.

William Morgan, Pi-Chuan Chang, Surabhi Gupta, and Jason M. Brenier. 2006. Automatically detecting action items in audio meeting recordings. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 96–103, Sydney, Australia.

Gabriel Murray and Steve Renals. 2007. Towards online speech summarization. In *Proceedings of INTERSPEECH 2007*, Antwerp, Belgium.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *Proceedings of the 10th European Conference on Speech Communication and Technology (INTERSPEECH - EUROSPEECH)*.

Matthew Purver, Patrick Ehlen, and John Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In *MLMI 2006, Revised Selected Papers*, Lecture Notes in Computer Science. Springer.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

Alison J. S. Sanford, Anthony J. Sanford, Jo Molle, and Catherine Emmott. 2006. Shallow processing and attention capture in written and spoken discourse. *Discourse Processes*, 42(2):109–130.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge, Massachusetts.

Kristina Toutanova and Christopher Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Who Tunes Accessibility of Referring Expressions in Task-Related Dialogue?

Ellen Gurman Bard
LEL,
University of Edinburgh,
Edinburgh EH8 9LL, UK
ellen@ling.ed.ac.uk

Robin Hill
HCRC,
University of Edinburgh,
Edinburgh EH8 9LW, UK
r.l.hill@ed.ac.uk

Mary Ellen Foster
Informatik VI,
Technische Universität,
München, Germany
foster@in.tum.de

Abstract

Ariel (1988; 1990; 2001) has proposed that the grammatical form of an anaphor can be predicted from the 'deemed' accessibility of its antecedent. The element of judgment in the term 'deemed' is critical: it allows the speaker to reflect an egocentric perspective and frees choice of expression from the actual contingencies of the situation in which it is uttered. Using a screen-based joint tangram construction task (Carletta et al., under revision), we examine the accessibility of 1775 introductory mentions for effects of situation (communication modalities and actions involving the named entity) and of responsibilities assigned to the participants. We find statistically significant effects of three kinds: circumstances readily available to the listener (concurrent movement of the named object); circumstances private to the speaker (hovering the mouse over the object, when the listener cannot see the mouse), and the speaker's role in the joint task. Since egocentrically selected forms may be underspecified, we make a preliminary attempt to discover whether referring expression usage is disabling or irrelevant.

1 Introduction

The question of what a thing shall be called has engaged psychologists and linguists as much as it engages anyone attempting automatic interpretation or generation of referring expressions (Brown, 1958; Dale & Reiter, 1995; Gundel, Hedberg, & Zacharski, 1993; Kranstedt, Lücking, Pfeiffer, Rieser, & Wachsmuth, 2006; Lyons, 1977; Prince, 1981; Van der Sluis & Krahmer, 2007; Walker & Prince, 1996). One very wide-ranging approach, (Ariel, 1988, 1990, 2001), attempts to key elaboration of the form of referring expressions to the 'deemed' accessibility of the referent, that is, to how difficult the producer of the expression estimates it will be to access the referent concept, discourse entity, or extra-linguistic object. Expressions introducing entities deemed completely unfamiliar to the audience should be maximally detailed indefinite NPs including modifiers of various kinds, as in (1). Expressions of intermediate accessibility might be marked by definite articles, deictic expressions, or personal pronouns in that order. Expressions making reference to a single most immediately mentioned entity in focus can be as minimal as so-called clitics (2), unstressed and all but deleted pronouns, or even zero forms (3).

(1) *A Republican governor of a strongly Democratic state*.
(2) A: Where's Arthur?
B: /z/ in the garage.
(3) A. And your younger son?
B. *{-}* playing Internet poker.

Accessibility theory provides a unified framework for predicting how forms of referring expressions will respond to givenness, discourse focus, inferrability from local scenarios and the like. As a general notion, accessibility ought to include effects of any available conditions which might draw attention to the correct referent, whatever modality delivers them and whether they are internal or external to the discourse. This paper discusses the accessibility of referring expressions produced during a joint construction task and examines two factors which might draw attention to the correct referent, task related movements and the roles of the participants.

The origins of our questions about these factors lie in the information which human interlocutors might use in determining how to refer. Ariel's notion of accessibility appears to depend on what the speaker supposes is the case, not on what is genuinely easier or more difficult for the

listener. While some approaches to dialogue assume that speakers carefully model their interlocutors, so that initial forms of expression could arise from the interlocutors' needs (Brennan & Clark, 1996; Clark & Krych, 2004; Schober, 1993), there is increasing evidence that we have limited ability to construct, recall, or deploy any such model in a timely fashion (Bard, Anderson et al., 2007; Horton & Gerrig, 2002, 2005a, 2005b; Horton & Keysar, 1996). Interlocutors may behave egocentrically (Bard et al., 2000; Bard & Aylett, 2004; Horton & Keysar, 1996), adopt a global account of affordances of a situation, (Anderson, Bard, Sotillo, Newlands, & Doherty-Sneddon, 1997; Brennan, Chen, Dickinson, Neider, & Zelinsky, In press), or observe information indicative of the listener's knowledge, but fail to act on it (Bard, Anderson et al., 2007; Brennan et al., In press).

The situation for form of referring expressions is mixed. While accessibility of referring expressions is more sensitive to the knowledge of the listener than is clarity of articulation (Bard & Aylett, 2004), other studies show that tendencies to match nomenclature to listener's history or current situation are quite variable (Brennan & Clark, 1996; Horton & Gerrig, 2002, 2005a; Horton & Keysar, 1996; Keysar, Lin, & Barr, 2003). So-called conceptual pacts are actually lexical pacts (Brennan & Clark, 1996), agreements to call objects by certain names, and are the result of negotiation over time, across which accessibility of the referring expression naturally rises. If speakers do track one another's internal states, the accessibility of even introductory mentions will suit the interlocutor's current needs, rather than the speaker's.

The evidence may be inconclusive because the typical paradigms for dialogue studies restrict cooperation to disjointed episodes. Often one participant instructs another to act on or select from an array of potential referents, while the other follows instructions relative to an identical or partially overlapping array. Both responsibilities and activities are clearly distinct. Even when players ultimately exchange roles, the roles are inherently asymmetrical: one has more information and more power to design the communication than the other. Channels for communication are purposely limited; and the knowledge shared between instructor and listener is altered trial by trial in an unpredictable way. To discover whether more robustly cooperative behaviour appears in more cooperative tasks, we have created a corpus of dialogues centred around a shared task which demands joint attention but makes it possible to vary the participants roles.

To study joint action as a model for human-robot cooperation in quasi-industrial settings, the JAST project has developed the Joint Construction Task (Carletta et al., under revision) in which two human players cooperate to construct a two-dimensional tangram on their yoked screens (Figure 1). Each player can manipulate the component parts by mouse actions. Each dyad is assigned to work either with roles (one player managing the task and the other assisting) or without. Mouse actions draw attention not just because they are integral to the construction process, but because, to mimic industrial risks, they are dangerous. If both mice touch the same object, or if two objects overlap, both break. Because each player can act on the tangram parts and sub-constructions, the activity of grasping or moving the named object adds a haptic or praxic modality to spoken forms. Even 'hovering' the mouse over a part without grasping it offers a chance to make a part accessible. The paradigm suggests how accessible initial mentions should be: because tangram parts come in identical pairs, a felicitous first mention should in theory be an indefinite expression like (4) or (5)

(4) Let's get *a red square*

(5) We could try *one pink triangle* first.

To discover how well keyed any change in form of referring expression is to the perceptions of the interlocutor, the design contrasts situations in which each player's mouse cursor is projected onto the other's screen with situations in which each player can see only the resulting movement of the object which the other's mouse 'grasps'. Only in the first case can a player see the mouse 'hovering' over a tangram part which is not actually moving.

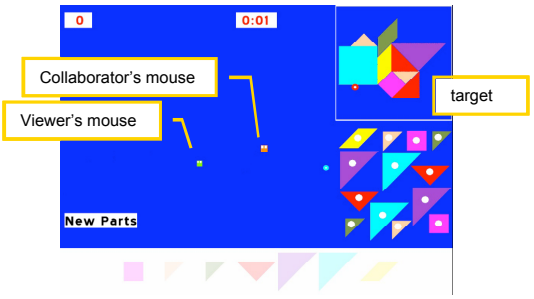


Figure 1. Joint Construction Task shared screen.

If moving a part draws attention, it should also give rise to referring expressions of greater accessibility. Since pointing is associated with shorter, less detailed referring expressions and pointing to closer targets has an even stronger effect (Kranstedt et al., 2006), touching and moving should have a very marked effect on the form of expression. Like Kranstedt et al., and exactly as the definition of deixis would predict (Lyons, 1977), we note the association of the 'hand' location and verbal deixis: in our case a larger proportion of verbal deictics (*this square; these, mine*) than of other forms of expression coincide with mouse-referent overlap (Foster et al., 2008).

To go further, we need to divide overlaps into those where the mouse is moving a part and those where it is merely hovering over it. If the listener's knowledge is of concern, a speaker moving parts and a speaker hovering over them should sometimes make different selections of accessibility level. Since movements of objects will always be visible to the listener in this paradigm, a speaker adjusting to listener knowledge could certainly use deictic forms to refer to parts she is currently moving. In contrast, visibility of the mouse cursor should determine whether a hovering mouse makes an object more accessible: Only when the hovering mouse cursor is visible to the listener can a speaker use it to point to the named object and select a more accessible referring expression. In fact, a speaker might even increase the accessibility of an expression referring to a part which the *listener* is visibly touching or moving. When the hovering mouse is not cross-projected, a listener-sensitive speaker cannot use it to point. If the listener's knowledge is less important than the speaker's, however, the speaker's own hovering movements should attract higher accessibility forms regardless of what the listener can see.

The players' roles suggest further questions. Managers have a primary role in setting the dyad's agenda. They should have more power to designate discourse focus and to change it, for example. If the choice of accessibility level is an overt designation on the speaker's part, then managers should have special powers of designation. Moreover, as we suggested earlier, managers might have less reason to track or adjust to the needs of their partners than role-less players do. Conversely, assistants should have more reason to adjust to the manager's precedents.

In all cases, the answers to our questions should be reflected in distributions of referring expressions across ordered levels of accessibility. Though accessibility bears on the relationships between earlier and later mentions of an entity, it ought to be important to determining the form of introductory mentions, too. By restricting our investigation in this way, and by controlling the objects available for naming, we can test our hypotheses about how a thing shall first be called.

2 Corpus Collection and Coding

2.1 Task

The Joint Construction Task or JCT (Carletta et al., under revision) offers to two collaborating players a target tangram (Figure 1, top right), geometrical shapes for reproducing it (centre right), a work area (centre screen), a counter for breakages (top left), a set of replacement parts (bottom of the screen), and a clock measuring elapsed time (top centre). The players' task is always to construct a replica of the target tangram as quickly, as accurately, and as cheaply in terms of breakages as possible. An accuracy score (top left) appears at the end of each trial.

Participants manipulate objects by left-clicking the mouse and dragging them or by right-clicking and rotating them. Carefully timed collaboration is required. Any part or partially constructed tangram 'held' by both players will break and must be replaced from the spare parts store to complete the trial. Moving an object across another breaks both. Objects can be joined only if each is held by a different player. Objects join permanently wherever they first meet. Inadequate constructions can be purposely broken and rebuilt from spare parts, incurring a cost in both parts and time.

Players' mouse cursors differ in colour and each changes colour when it has grabbed an object, distinguishing grabbing from mere superimposition (hovering).

2.2 Apparatus

Each participant sat approximately 40cm from a separate CRT display in the same sound-attenuated room. Participants faced each other, but direct eye contact was blocked by the monitors between them. Participants were eye-tracked monocularly via two SR-Research EyeLink II head-mounted eye-trackers. Head worn microphones captured speech on individual channels. Continuous audio and video records included a full account of locations and movements of indi-

vidual parts, constructed objects, and cursors. Composite videos recorded all movements and audio.

2.3 Participants, design and materials

Sixty-four Edinburgh University students, paid to participate, were paired into 32 same-sex dyads who had never met before. Four further dyads were discarded because of technical failures. Each dyad participated in 8 experimental conditions produced by the factorial manipulation of three communication modalities: speech, gaze (each player's current eye-track cross-projected onto the other's screen), and mouse cursor (also cross-projected). Participants could always see their own mouse cursor. Without no additional communicative modalities, they saw only the moving parts. Gaze and mouse conditions were pseudo-randomised following a latin square. Speech and non-speech conditions were counterbalanced. Only conditions with speech are analyzed here.

In 16 dyads, one participant was designated manager and the other assistant. The manager was instructed to maintain speed, accuracy, and cost, and to signal the completion of each trial. The assistant was to help. The remaining dyads were assigned no roles but otherwise had the same working instructions. Trials ended when one player declared the construction complete by pressing the spacebar and the other confirmed. An accuracy score reflecting similarity between the built and the target tangrams then appeared across the built exemplar.

Each dyad reproduced 16 different tangrams, 2 per condition. No tangram resembled a nameable object. Each contained 11 parts. All trials used the same set of 13 parts, comprising 2 copies of each of 6 shape-colour combinations (squares or right-angle isosceles triangles differing in size and colour) and a single yellow parallelogram. These initially appeared in 4 different layouts counterbalanced across experimental items. The extra pieces differed from trial to trial.

2.4 Coding referring expressions

Dialogues were transcribed orthographically. Each referring expression was time-stamped for start and end points. Then each expression referring to any on-screen object was coded with a referent identifier linking it to the object. Coders had access to the video and audio track and were allowed to use any material within a dialogue to determine the referent of any expression. All referring expressions were coded for accessibility on the scale given in Table 1. This system represents a modest expansion of a system applied to an earlier corpus of task-related dialogues (Bard & Aylett, 2004) and yielding negligible disagreement between coders.

Table 1 Accessibility Coding Scheme

Level	Definition	Examples
Min	Indefinite NP	*a purple one* *one of the nearest blue pieces*
	Bare nominal	*pink one* *triangles*
	Definite NP	*the red bit* *the other purple one*
	Deictic NP	*those two little kids.*
	Deictic / Possessive } Pron	*these* *mine*
Max	Other Pronouns	*it*
	Clitic/inaudible.	*-/z/*

3 Results

3.1 Overall outcome

Figure 2 presents the overall distribution of first mentions across the accessibility scale. Despite the fact that most original parts would be expected to demand an indefinite referring expression to distinguish them from an identical part, only 16% of first mentions were indefinite NPs. The remaining 84% were of higher accessibility. Our question now is whether mouse actions or speaker roles are responsible for this profile.

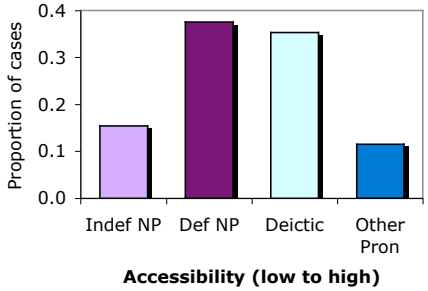


Figure 2. Accessibility of first mentions in the Joint Construction Task

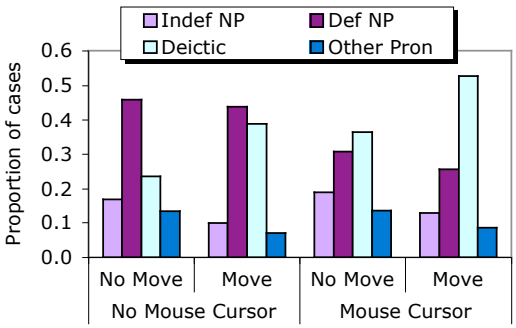


Figure 3. Accessibility of first mentions: Effects of moving the referent object

3.2 Modalities, roles, and accessibility

Method. The conditions critical to our predictions were coded for a multinomial logistic regression which modelled the distribution of first mentions across accessibility categories. This statistic tests the capacity of category variables (like Mouse v No Mouse) to influence ordinal variables (like Accessibility). It constructs regression equations both for the whole ordinal series and for the comparison of each level to some reference level. We use it to ask which actions and modalities change the tendency to produce indefinite referring expressions (the usually expected format) relative to each more accessible category.

The calculations are done on log odds, but for interpretability, we display simple proportions of cases. To reduce the number of empty cells, accessibility categories were collapsed into four levels: Indefinite NPs (including bare nominals), Definite NPs, deictics (including deictic NPs, deictic pronouns and possessive pronouns) and other pronouns (including clitics).

Separate equations were prepared for the Mouse Cursor Cross-Projected (n = 836) and No Mouse Cursor conditions (n = 939). The predictors included the experimental variable Roles Assigned, the participants' mouse actions (the speaker/listener moving part being mentioned, or 'hovering' the mouse over it), and the interactions of Roles Assigned with each movement variable. Gaze cross-projection was not included, as it had proved an ineffective predictor in earlier exploratory regressions. Table 2 shows the significant outcomes. Each effect listed is essentially independent of any effect from any concurrent predictor.

No effect of listener behaviour reached significance. There were effects of the speaker's actions and of Role Assignment.

Table 2. Significant predictors of accessibility. For individual levels of accessibility, $df = 1$. $* = p < .05$; ‡ = $p < .01$; § = $p < .001$.

No Mouse Cursor Cross-Projection						
	-2 Log Likelihood		χ^2	*df*	Cox & Snell	
	268.07		105.00§	27	0.106	
	Speaker Move		Speaker Hover		Speaker Hover x Roles Assigned	
χ^2					276.00	7.42*
	B	Wald	B	Wald	B	Wald
Definites			-1.137	10.85§	1.075	4.99*
Deictics	-0.814	4.78*			1.275	6.17*

Mouse Cursor Cross-Projected						
	-2 Log Likelihood		χ^2	*df*	Cox & Snell	
	258.00		61.34§	27	0.071	
	Speaker Move		Speaker Hover		Speaker Hover x Roles Assigned	
χ^2					266.00	7.77*
	B	Wald	B	Wald	B	Wald
Deictics	-0.722	5.05*				
Pronouns			1.264	6.95‡		

As predicted, actions available to speaker and listener were important: visibly moving the referent (Figure 3) coincided with increased deictic expressions (31% without v 46% with movement overall) at the expense of indefinites (18% v 12% overall) whether or not the speaker's mouse cursor itself was visible. Also, visibly hovering the mouse cursor over the referent (Figure 4) accompanied a significant fall in pronouns (15% v 6%) relative to indefinites (17% v 16%), with deictics the dominant category in both cases (39%, 51%).

Strikingly, actions unavailable to the listener were also important. An *in*visibly hovering mouse accompanied a shift from indefinites (17% v 10%) towards definites (42% v 50%), with the latter as the most common category.

Role assignment influenced the effects of invisible mouse gestures: Only in Manager-Assistant dialogues did introductory mentions shift markedly away from indefinites (22% v 8%) toward deictics (25% v 40%) as well as definites (36% v 48%). Figure 4 shows that in Manager-Assistant dialogues private hovering gestures gave profiles somewhere between their No Roles counterparts (where definite NPs predominate) and dialogues with projected cursors (where deictics rise with public gestures).

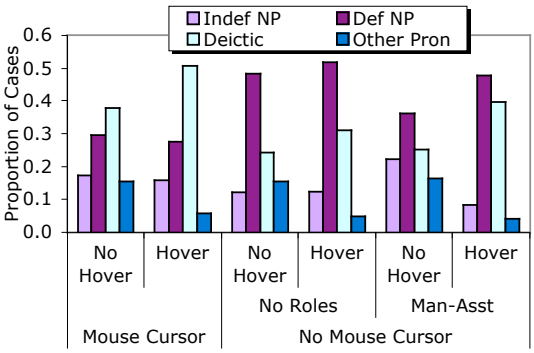


Figure 4. Effects on accessibility of hovering mouse over referent, by cursor visibility and assigned roles.

4 Discussion

This paper asked whether the association between handling a thing and using an accessible format to name it was linked to the speaker's own knowledge or to the knowledge expected to reside with the listener. There are two reasons to believe that the listener is not in charge. First, we found no significant effects of the listener's manipulation of tangram parts on the speaker's form of referring expression, even when the listener's movements were fully visible to the speaker. Second, we did find effects of speakers' mouse gestures which were invisible to the listener.

At the same time, we suggested that if accessibility is an expression of opinion, it should be manipulated by Managers in particular. In the event, Manager-Assistant dialogues showed more egocentric use of accessibility than no-role dialogues: in these dialogues the presence of a gesture invisible to interlocutor all but eliminated indefinite introductory mentions, in favour of definites and deictics.

While the effect of movement is a praxic or haptic form of deixis, the effects of private gestures and of role ought to be counterproductive. Though all the conditions examined here yield tangrams of equal similarity to their models, the costs do follow this prediction. Trials without mouse cross-projection took longer than those with it (205 v 187sec; $F_1 = 11.45$, $df = 1, 30$, $p = .002$) and incurred more breakages (1.8 v 2.3: $F_1 = 4.52$, $df = 1, 30$, $p = .008$) to achieve equal accuracy (92.1 v 91.9). Manager-Assistant trials took longer than No Role trials (216 v 175sec: $F_1 = 10.67$, $df = 1, 30$, $p = .003$) to give similar performance (Accuracy: 93.8 v 91.2; Breakages: 2.0 v 2.1). The latter finding is the stronger argument, because additional breakages require additional time to fix.

Nonetheless, the picture is far from complete. We see three major issues.

First, the results fall some way short of a clear case for managerial insensitivity. The Role Assignment results were based on expressions produced by both participants. Analyses comparing managers with assistants are made difficult by small or empty cells. Both players show the pattern found in Figure 4. Accordingly, we have no particular evidence contrasting managers with their assistants, though we can distinguish manager-assistant dyads from the dyads who had no assigned roles.

As we suggested earlier, however, one result of role differences is to give precedence to one individual. The manager decided what should happen next. To cooperate, the assistant had to conform to the manager's choices. Conforming to the manager's referential habits, for social reasons, or through structural priming, could make the assistant appear to designate with invisible gestures, too. In essence, the assistant can achieve a tendency toward use of definites or deictics where they might not otherwise appear to be unwarranted and then employ private gestures to

accompany these instances. In contrast, No Role dyads might follow a mixture of styles or compete to control the task plan or the naming habits. If so, manager and assistant should have more similar profiles in than No Role players. Quantitatively, this seems to be the case.

Second, though it is clear that speakers' private and public actions associate with particular levels of accessibility, it is not clear that their effects are all increases in accessibility. For example, Figures 3 and 4 show a tendency, significant only with hovering, for speaker actions not to collocate with the highest levels of accessibility in first mentions: pronominal or clitic introductory mentions are used less often when the mouse overlaps the referent part than when it does not. Thus, the haptic or ostensive functions of mouse movements are specific to definite and deictic usage: they literally turn *a triangle* into *this* but they do not turn *this* into *it*. For this reason, the single accessibility continuum might be viewed as the result of a set of different referential phenomena, for example, demonstration or givenness in context, bearing on speakers' choices with different degrees of force.

Finally, there is the issue of the discourse history within which the introductory mentions are set. Clearly, some first mentions do not refer to totally discourse-new or completely unpredictable entities (Prince, 1981). There is no doubt that other forces work on the choice of referring expressions.

We do not yet know how the sequence of external events – construction of the tangram, for example, affects the forms of introductory referring expressions. In theory, it is possible that corpus dialogues went well because speakers were attending to the same objects regardless of the form of referring expressions. Our eye-tracking results from this corpus suggest, however, that alignment between players was far from perfect. If players were already attending to the same objects, whatever either said, we should find very high levels of overlapping gaze. To study the relationship between interlocutors' gaze patterns, we have used a cross-recurrence analysis (Richardson, Dale, & Kirkham, 2007), which shows how behaviours can be entrained even if they are not synchronized to the extremely fine levels that eyetrackers detect. This technique shows what percentage of interlocutors' gaze fixations are on the same objects but separated by various lags in one direction or another. Almost without exception ((Bard, Hill, Nicol, & Carletta, 2007), maximal shared view was simultaneous, but it was far from complete: participants showed a maximum of 36% gaze at the same objects in conditions with speech (as against 40% without). It would seem that referring expressions still have some work to do when joint attention is required.

We began by discussing referring expressions in the light of speakers' ability to maintain models of their listeners' knowledge that update quickly enough to be the basis of initial mentions. Our speakers did not appear to use such models. Ultimately, of course, they could have had ample opportunity to produce adequate reference by subsequent joint adjustment. The additional time taken for dialogues in the more egocentric conditions suggests that this could be the case.

If so, the speakers' behaviour is another example of joint responsibility for dialogue being effectively shared rather than duplicated across interlocutors (Bard, Anderson et al., 2007). In this kind of responsibility structure, rather than a fully articulated model of common ground, a simple and risky egocentric process guides production. Speakers were free to designate invisibly just because their partners were under an obligation to object to inadequate expressions.

Acknowledgments

This work was funded by EU Project JAST (FP6-003747-IP). The authors are grateful to the JCT programmers, Tim Taylor, Craig Nicol, Joe Eddy and Jonathan Kilgour for their contributions to the software, to Jean Carletta who managed the program development, to the reference coders for guessing what the speaker had in mind, and to JP de Ruiter for helpful discussions.

References

Anderson, A. H., Bard, E. G., Sotillo, C., Newlands, A., & Doherty-Sneddon, G. (1997). Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception and Psychophysics, 59*(4), 580-592.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics, 24*, 65-87.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London.: Routledge/Croom Helm.

Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schilperoord & W. Spooren (Eds.), *Text representation:Linguistic and psycholinguistic aspects.* (pp. 29-87). Amsterdam: John Benjamins.

Bard, E. G., Anderson, A. H., Chen, Y., Nicholson, H. B. M., Havard, C., & Dalzel-Job, S. (2007). Let's you do that: Sharing the cognitive

burdens of dialogue. *Journal of Memory and Language, 57*(4), 616-641.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42*, 1-22.

Bard, E. G., & Aylett, M. P. (2004). Referential form, word duration, and modelling the listener in spoken dialogue. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*. (pp. 173-191). Cambridge, MA: MIT Press.

Bard, E. G., Hill, R., Nicol, C., & Carletta, J. (2007). *Look here: Does dialogue align gaze in dynamic joint action?* Paper presented at the AMLaP2007, Turku, Finland.

Brennan, S. E., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (In press). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 1482-1493.

Brown, R. (1958). How shall a thing be called. *Psychological Review, 65*, 14-21.

Carletta, J., Nicol, C., Taylor, T., Hill, R., de Ruiter, J. P., & Bard, E. G. (under revision). Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods, Instruments, and Computers*.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*, 62–68.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science, 19*(2), 233-263.

Foster, M. E., Bard, E. G., Guhe, M., Hill, R., Oberlander, J., & Knoll, A. (2008). *The roles of haptic-ostensive referring expressions in cooperative task-based human-robot dialogue.* Paper presented at the Human Robot Interaction, Amsterdam.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*, 274-230.

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language, 47*(4), 589-606.

Horton, W. S., & Gerrig, R. J. (2005a). Conversational common ground and memory processes in language production. *Discourse Processes, 40*(1), 1-35.

Horton, W. S., & Gerrig, R. J. (2005b). The impact of memory demands on audience design during language production. *Cognition, 96*(2), 127-142.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25–41.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006). Deictic Object Reference in Task-oriented Dialogue. In G. Rickheit & I. Wachsmuth (Eds.), *Situated Communication,* (pp. 155-207). Berlin: Mouton de Gruyter.

Lyons, J. (1977). *Semantics* (Vol. 2): Cambridge, UK.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics* (pp. 223-256). New York: Academic Press.

Richardson, D. C., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*(5), 407-413.

Schober, M. (1993). Spatial perspective-taking in conversation. *Cognition, 47*(1), 1-24.

Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal referring expressions. *Discourse Processes, 44*(3), 145-174.

Walker, M. A., & Prince, E. F. (1996). A bilateral approach to Givenness: A hearer-status algorithm and a centering algorithm. In T. Freitheim & J. Gundel (Eds.), *Reference and referent accessibility*. (pp. 291-306). Amsterdan: John Benjamins.

What's in a manner of speaking?
Children's sensitivity to partner-specific referential precedents.

Danielle Matthews
Max Planck Child Study Centre
School of Psychological Sciences
University of Manchester, U.K.
danielle.matthews@manchester.ac.uk

Elena Lieven & Michael Tomasello
Max Planck Institute
for Evolutionary Anthropology
Leipzig, Germany
[lieven][tomas]@eva.mpg.de

Abstract

This study investigated whether young children form 'referential pacts' (Brennan & Clark, 1996; Metzing & Brennan, 2003) such that they expect people to refer to objects with the same terms over time unless there is a good reason to switch to using a new expression. 128 children aged 3 and 5 years participated in a study where they cooperated with an experimenter (E1) to move toys around to new locations on a shelf. E1 established referential terms for all the toys in a warm up game. Then, either E1 (**original partner condition**) or a new experimenter, E2 (**new partner condition**), played a second game with the same toys. In the second game, two critical toys were referred to with their **original terms** and two with **new terms**. Children were significantly slower to pick up a toy if it was referred to with a new term than with an old term. Crucially, this difference in reaction times was significantly greater in the original partner condition. This suggests that children found it harder to process a new term when it was produced by someone who had previously referred to the same toy with a different expression. That is, children as young a 3 years of age show adult-like sensitivity to referential pacts.

1 Introduction

According to Grice's Maxim of Manner, speakers should not abandon a perspective without good reason. So, if we are engaged in moving some toys around on a set of shelves and I refer to a toy consistently as '***the bush***', then you will come to expect me to continue to use that term to refer to the same object in the future. If I suddenly abandon our 'referential pact' and call the toy '***the tree***' you will be momentarily confused. However, if a new person (with no prior experience of our pact) enters the room and uses the alternative referring expression ('***the tree***'), you would not find it confusing, as long as it is an acceptable description of the toy in the absence of a prior pact (Brennan & Clark, 1996; Metzing & Brennan, 2003).

In an experimental investigation of adult sensitivity to referential pacts, Metzing and Brennan (2003) had participants play a cooperative game of the type described above with an experimenter who established shared terms for objects (e.g., ''the shiny cylinder'') during repeated references to them. After this warm up phase, either the original experimenter or a new experimenter (who had not observed the warm up) continued the game and used either the original expressions or a new ones (e.g., ''the silver pipe'') to refer to the previously discussed objects. In this test phase, adults were equally quick to comprehend original expressions regardless of which experimenter produced them. However, when objects were referred to with new expressions, there was partner specific interference: adults were 12 milliseconds slower to touch the target object when the new expression

was uttered by the original experimenter than when the new expression was uttered by the new experimenter. This difference in reaction times was also reflected in the adults' eye movements to target objects and was argued to reflect adult sensitivity to referential pacts – if someone suddenly switches from using one term to using another for no apparent reason, it slows you down.

Metzing and Brennan's (2003) finding that comprehension of referential terms is subject to partner-specific effects is now generally accepted. However, debate continues as to how early this effect of referential pacts is in adult processing (Brown-Schmidt, 2008; Kronmüller & Barr, 2007). There is also controversy concerning whether referential pacts rest on a principle of co-cooperativeness that is mutually assumed to hold between two conversational partners or whether pacts are a reflection of a more simple expectation that people will be consistent in their use of expressions across time (Shintel & Keysar, 2007).

Whatever the outcome of the above debates, it is unclear when we would expect children to show sensitivity to referential pacts. It is plausible that before the age of four, children would expect everyone to use the same term for an object regardless of whether they were present when a pact was established. Indeed, studies on the development of synonyms suggest that three-year-olds will not accept that a given toy can be called, for example, both 'a rabbit' and 'a bunny' (Doherty, 2000; Doherty & Perner, 1998; Perner, Stummer, Sprung, & Doherty, 2002). In these alternative naming studies, children aged between three and five years were instructed that if a puppet calls an item, e.g., 'a rabbit' the child has to call it something else, e.g., 'a bunny' or, in a judgment version of the task, the child has to name a toy and then, when the puppet refers to it with an alternative term, the child has to say whether the term is acceptable or not. The synonyms used in this task were: bunny–rabbit, lady–woman, television–TV, coat-jacket. In control games, the children were asked to name a colour of the item (e.g. Puppet: "bunny", Child: "white") or part of the item (Puppet: "bunny", Child: "tail"). Three-year-olds tended to fail the alternative name task (insist that a bunny cannot also be called a rabbit) despite passing the control task, whereas older children tended to pass the alternative naming task at around the same time they began to pass false belief tasks. The explanation of these results was thus that before four years of age children cannot reconcile conflicting perspectives in order to understand that what one person might call a bunny another might call a rabbit. Although not adult like, this kind of mutual exclusivity constraint has been argued to convey certain advantages in early language learning (c.f. Sabbagh & Henderson, 2007).

Given the above findings we were interested to investigate whether young children are sensitive to referential pacts and whether this sensitivity only emerges after 4 years. We thus adapted Metzing & Brennan's (2003) task for use with children. In a within-subjects design, children played with two sets of toys. With one set, experimenter 1 (E1) established names for the toys in a warm up phase and then continued to play in the test phase. With the other set of toys, E1 played the warm up phase and then a new person, E2, played the test phase. Each test phase had four critical toys: toys 1 and 3 were referred to with an original expression established in the warm up phase and toys 2 and 4 were referred to with an entirely new expression. We recorded how long it took children to pick up each toy. Thus for each test phase we were able to make two comparisons: whether children were quicker to pick up toy 1 than toy 2 (trial 1) and whether they were quicker to pick up toy 3 than toy 4 (trial 2). Of greatest interest is whether any differences in reaction times vary as a function of the identity of the experimenter.

2 Method

2.1 Participants

126 normally developing, monolingual, English-speaking children were included in the study (51 boys, 75 girls). There were 62 three-year-olds (range 3;0-3;11, mean age 3;5) and 64 five-year-olds (range 5;0-5;11, mean age 5;6). The children were tested in a university laboratory in the U.K.. Full parental consent was obtained for each child.

2.2 Materials and Design

Fourteen toys were selected on the basis that they could be described felicitously by two different, well-known nouns that occur frequently in the speech directed to 3-year-old children (as verified by a search of the CHILDES database, MacWhinney, 2000). Of these fourteen, eight were

selected as stimuli on the basis that a group of 16 3-year-olds (not tested in any of the subsequent procedures) used at least two different well-known words to spontaneously refer to the toys when asked '***What's this?***'. These preferred terms were then used as the referring expressions for the study. The pairs of terms used to describe the 8 critical toys are presented in table 1. One set of toys was used for the 'same partner' condition and another set for the 'new partner' condition (counterbalanced).

Table 1. Pairs of referring expressions used to refer to critical toys.

Set A	**Set B**
car / truck	girl / lady
book / story	pillow / cushion
horse / pony	turtle / tortoise
tree / bush	nose / apple

To be confident that most 3-year-olds would be able to identify each toy upon hearing either of the above terms, we first conducted a comprehension test with two groups each made up of seven 2-year-olds and 12 3-year-olds. Again, none of these children took part in the main study. Both groups saw all the toys at the same time and were asked to '***find the*** [toy name]'. The first group heard the first of the alternative terms (***car, nose, book*** etc.) and the second group heard the second of the terms (***bush, cushion, nose*** etc.). In all cases at least five 2-year-olds and 11 3-year-olds were able to identify each toy on the basis of the terms they heard.

For each partner condition in the main study, we put one set of test toys along with 8 'filler' toys into a 5 x 3 block of Perspex pigeonholes (see figure 1). The arrangement of the toys was fixed such that an experimenter could instruct the child to rearrange them following a script. Photographs of each set of toys in differing arrangements were taken and used as props, as explained below. Figure 1 a and b present example arrangement for both of the set of toys. A video camera was set up at the edge of the Perspex boxes such that it was possible to code at precisely which frame the child's hand entered a box to retrieve a toy (see figure 2). Two other video cameras recorded the child and the experimenter as they interacted.

2.3 Procedure

Upon arrival, the child and their care-giver(s) entered the test room and the child was allowed to play freely with E1 while E2 obtained parental consent for the study. This ensured the child had seen both experimenters before the test began.

Figure 1. Toy sets A and B

After a period of free play, E2 left the room and the child sat with E1 at a table in front of the Perspex boxes that were covered over with a piece of cloth to prevent the child from spontaneously naming the objects. E1 explained that under the cover there were lots of toys and that she had a photo of where the toys should go. E1 showed the child the first photograph briefly at this point. She then suggested that she could look at the picture to see what needed moving round and the child could find the toys and put them in the right places. She asked the child if s/he would like to help and when the child agreed E1 said that they would manage to do it together. E1 then lifted the cover to reveal the toys for the first game.

Each child played four 'games', two per condition. Each game consisted of rearranging the toys so that they matched a photograph. The first game of each condition served as a warm-up in

which all the key referring expressions were introduced and entrained upon. This first game was always played with E1. It consisted of a sequence of 16 instructions of the basic form '***Get the X, put it next to/under/above the Y***'. Hesitations and hedges (e.g. '***Now get…I think it's Lego….can you see any? Yes, put it under the…er…man'***) were written into the script to reinforce the impression that the experimenter didn't have a fixed conceptualization of all of the toys from the outset. Each of the 4 critical test objects was referred to 4 times.

For accuracy of coding, it was important to ensure that children's hands were always in the same position on the table before they took an object out of a box. To achieve this, after 12 warm up instructions E1 showed the child a pair of red hands that had been drawn on the table and asked the child to put his/her hands on the red hands before they began each turn. From this point on, E1 ensured that the child returned their hands to the red hands on the table before each new instruction ' to show they were ready'.

Once all the warm-up instructions had been carried out, E1 announced that the toys looked the same as in the photo. She showed the child the photo to see if s/he agreed and remarked on what a great job they had done. E1 let the child chose a sticker as a reward and asked if s/he'd like to play another game. E1 then left the room on the pretext of needing to get another photo to make. She returned after a minute and suggested that they make the next photo. At this point E2 entered the room and explained that the secretary needed E1, asking if she could come and help her for a minute. E1 protested that she was just needed to play a game quickly and asked if she could come in a minute. What happened next varied according to the two experimental conditions.

In the **same speaker condition**, E2 acquiesced and said she would explain to the secretary that E1 would come in a minute. E1 then played the second game of that condition with the child. In the **new speaker condition**, E2 told E1 that the secretary really needed her help now. E1 agreed to go, asking E2 if she could quickly play the game with the child. E2 said she was not sure what to do but E1 reassured her it was easy and said '***You just need to make this look the same as my picture so you need to move the toys around. Like you might say "get the [filler item] and put it next to the [filler item]". . CHILD'S NAME will help you. We always put our hands on the red hands before we start to show we are ready. I'll be back in a minute.***'. E1 left the room and E2 played the second game with the child remarking that it didn't look too difficult and that she hadn't seen the toys before.

The second game consisted of 7 scripted instructions and was played in the same manner as the first, ensuring the child's hands were always on the red hand markers before beginning the next instruction. Instructions 1, 2 and 7 referred only to filler toys. Instructions 3 and 5 referred to two of the critical toys with the **same expressions** as had been previously used in the warm-up game. Instructions 4 and 6 referred to the other two toys with **different expression**s to the ones used in the warm up. Instructions 3,4,5, and 6 are henceforth referred to as critical trials with instructions 3 and 4 being referred to as trial 1 and instructions 5 and 6 being referred to as trial 2.

Half the children took part in the same speaker condition followed by the different speaker condition. The other half had the opposite order. Whichever condition came first always used toy set 1 and the second condition always used toy set 2. Scripts were fully counterbalanced so that, for each pair of referring expressions both terms were heard equally often as a) the same expression used twice (e.g. warm-up game: '***Tree'***, test game '***Tree'***), b) the first expression before a switch (e.g. warm-up game: '***Tree'***, test game '***Bush'***), c) the new expression after a switch (e.g. warm-up game: '***Bush'***, test game '***Tree'***). All the scripts were written so that the critical toys would be on the middle row before the test game began. This ensured the children would have the same distance to reach each toy. Furthermore, the scripts and accompanying photographs were counterbalanced so that the critical toys appeared on the shelf in two different orders from left to right. This ensured that if any of the boxes was in a privileged position on the shelf (i.e. that was quicker to reach) it would not affect the reaction times for a given condition. Finally, the identity of experimenter 1 and 2 was fully counterbalanced. The same full-time research assistant performed the role of experimenter 1 for half the children in each age group and experimenter 2 for the other half. The other experimenter role was performed by one of three other assistants.

2.4 Coding

The videos of the children's hand movements when retrieving toys were coded using Adobe Premier software. A research assistant coded the length of time it took from the onset of the critical referring expression (as located on the audio wave) for the child to reach into the relevant box (the first frame where the fingertips were inside the box). Very rarely, children retrieved and object that was not the target. These cases were excluded from analysis.

3 Results

Table 2 reports the reaction times for both ages and trials as a function of partner identity and referential term.

Table 2. Reaction times in seconds.

		Trial 1		Trial 2	
		Same Term	New Term	Same Term	New Term
3yrs	Same Partner	2.7	4.4	2.8	3.8
	New Partner	2.4	3.0	2.8	3.6
5yrs	Same Partner	1.8	2.9	2.3	3.1
	New Partner	1.8	2.3	2.1	2.7

To facilitate statistical analysis we converted these raw reaction times to difference scores (RT to New term – RT to Original term). These difference scores are presented in figure 2.

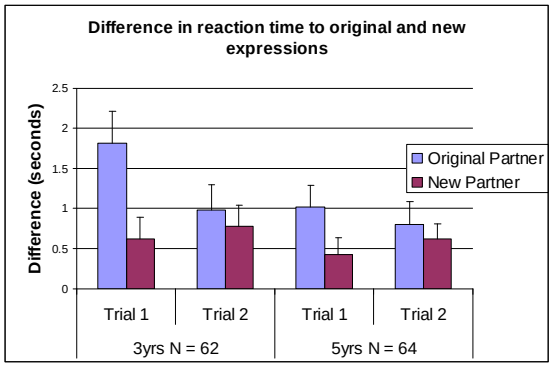


Figure 2. Difference in reaction time (new expressions minus original expressions)

Wilcoxon tests confirmed that on the first trial children were slowed down by the use of a new referring expression significantly more if the expression was produced by the original partner than if by a new partner ($Z = 2.561$, $\boldsymbol{p} = .01$). This effect was more pronounced in the younger children and indeed when each age is considered separately only the effect of partner identity is only significant for the three-year-olds, ($Z = 2.068$, $\boldsymbol{p} = .039$). There were no significant effects for trial 2, which would suggest that once a pact has been broken 'all bets are off': children are not surprised if subsequently other pacts are also not adhered to.

To investigate whether the effects observed in trial 1 were carried by particular items, we fitted a mixed effect regression model to the data with child, new term and original term (for each object) as random variables, age, partner identity and the interaction between these two factors as fixed effects and difference in reaction times on trial 1 as the outcome variable (Baayen, 2008). Partner identity was a significant predictor ($\boldsymbol{B} =$ 1.9931, $\boldsymbol{p} =$ 0.0344) such that difference scores were greater in the original partner condition. Age and the interaction between age and partner identity were not significant predictors.

4 Discussion

These results suggest that children show sensitivity to referential pacts from a young age. Like adults, children found it harder to process a new term for an object if it was produced by someone who had previously referred to the same object with a different expression. Interestingly this effect was only observed for the first trial of each test phase. This suggests that once someone has broken their 'referential history' children no longer expected them to adhere to it for subsequent reference to other objects.

From a developmental point of view, the current findings are surprising given that the three-year-olds we tested would not be expected to pass other tasks that require an understanding that whereas one person might call an object 'a tree' another might quite legitimately call it 'a bush'. Children were generally capable of processing two different terms for an object and were only slowed down in the comprehension of alternative terms by about 1 second - so long as their conversational partner was not breaking a referential pact. This would suggest that, at least in some circumstances, children are relatively flexible in understanding that an object may be referred to in different ways

by different people (Deák & Maratsos, 1998). Occasionally, some children were incapable of identifying an object given the new term of reference (and were accordingly given a maximum RT of 10 seconds, after which time the experimenter pointed to the target object). Thus on occasion, children were truly incapable of accepting two descriptions for one object. What the current results indicate, however, is that these cases are the exception rather than the rule.

Despite their ability to comprehend two different terms for one object, many children indicated that they were not happy with the use of the new term. They would often protest, saying, for example, 'It's not a tree, it's a bush!'. These protests were commonplace and indicate on the one hand that children detect a difference in perspectives about the same object, but on the other that they do not approve of it. Thus is would appear that children are 'hyper-conventional' at an early age. At the same time as understanding that the alternative terms where intended for the same object, they are very keen to pass normative judgment on their use. Children always preferred that the original term be maintained. Given the counterbalanced design, this suggests that children's protests were not based on their general preference for one term over another but rather based on a preference they created during the warm up trial.

With respect to the debates in the adult literature, the current results are informative to the extent that they demonstrate that referential pacts are not a highly controlled phenomenon that only adults would be capable of displaying. Whatever the preferred explanation of referential pacts - be they truly co-operative in nature or more expectation based – it is clear that they have an effect from early on in development and indeed are more pronounced for younger children. Apparently the older children were able to recover from a 'broken pact' faster than their younger counterparts. It would therefore not be surprising if such effects went undetected in adults at least some of the time, given how quickly they can be resolved in highly constrained contexts.

References

Baayen, H. (2008). ***Analyzing linguistic data: A practical introduction to statistics.*** Cambridge: Cambridge University 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.

Brown-Schmidt, S. (2008). Time course of processing conceptual pacts in conversation reveals early partner specific effects. ***Poster presented at CUNY 2008 Conference on Human Sentence Processing***.

Deák, G., & Maratsos, M. (1998). On having complex representations of things: Preschoolers use multiple words for objects and people. . ***Developmental Psychology, 34***(2), 224-240.

Doherty, M. J. (2000). Children's understanding of homonymy: metalinguistic awareness and false belief. ***Journal of Child Language, 27***(2), 367-392.

Doherty, M. J., & Perner, J. (1998). Metalinguistic awarenness and theory of mind: Just two words for the same thing? ***Cognitive Development, 13***, 279-305.

Kronmüller, E., & Barr, D. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preeemption hypothesis. ***Journal of Memory & Language, 56***(3), 436-455.

MacWhinney, B. (2000). ***The CHILDES project: Tools for analyzing talk*** (Vol. 2: The database). London: Lawrence Erlbaum Associates.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. ***Journal of Memory & Language, 49***, 201-213.

Perner, J., Stummer, S., Sprung, M., & Doherty, M. (2002). Theory of mind finds its Piagetian perspective: why alternative naming comes with understanding belief. ***Cognitive Development, 17***, 1451-1472.

Sabbagh, M., & Henderson, A. (2007). How an appreciation of conventionality shapes early word learning. ***New Directions for Child and Adolescent Development, 115***, 25-37.

Shintel, H., & Keysar, B. (2007). You said it before and you'll say it again: expectations of consistency in communication. ***Journal of Experimental Psychology: Learning Memory, and Cognition, 33***(2), 357-369.

Dimensions of Variation in Disfluency Production in Discourse

Scott H. Fraundorf
University of Illinois at Urbana-Champaign
603 E Daniel St.
Champaign, IL 61820 USA
sfraund2@uiuc.edu

Duane G. Watson
University of Illinois at Urbana-Champaign
603 E Daniel St.
Champaign, IL 61820 USA
dgwatson@uiuc.edu

Abstract

This study demonstrates that four common types of disfluency in discourse (fillers, silent pauses, repairs, and repeated words) differ from one another on two dimensions related to language production processes: their temporal relation to speech production problems and the level of production at which those problems occurred. Participants' speech in a story-telling paradigm was coded for the four disfluency types. Comparisons between types in their relation to story events, to clause boundaries, to utterance length, to utterance position, and to other disfluencies suggest the four types reflect different difficulties in language production. Temporally, fillers, silent pauses, and repeats represent difficulties in upcoming speech, while repairs represent past difficulties. Fillers were most associated with discourse-level problems, while silent pauses were more associated with grammatical and phonological difficulty.

1 Introduction

Human speech is fraught with interruptions, or *disfluencies*. Although several types of disfluency occur in speech, the ways in which these types differ from one another have not been well defined. In this paper, we propose that disfluency types systematically differ along at least two dimensions: (a) their temporal relationship to the underlying production difficulty and (b) the level of production at which the difficulty occurs.

Precise taxonomies of disfluencies vary, but most are derived from the four categories proposed by Maclay and Osgood (1959). *Fillers*, as in (1) below, are verbal interruptions that do not relate to the proposition of the main message—in English, most commonly *uh* and *um*. *Silent pauses*, as in (2), are periods of silence longer than the pauses that would be produced in an equivalent fluent utterance. *Repeats*, as in (3), are unmodified repetitions of a word or of a string of words. Finally, *repairs* are self-corrections or revisions of material already spoken. Repairs such as (4), called *error repairs* by Levelt (1983), simply correct errors of linguistic form. Repairs like (5), called *appropriateness repairs* and *message repairs* by Levelt, present a new or rephrased message.

(1) She grabs the fan and **uh** one pair of gloves.
(2) She sees **...** a small **...** box saying "EAT ME."
(3) Alice doesn't think **that cats that cats** grin.
(4) The cake **make Alices makes Alice** grow.
(5) And they sent Bill the lizard down the chimney **to find her er to see what was going on**.

Because these four types of disfluencies obey different distributional patterns and the frequency of use of each type correlates only weakly with that of other types, Maclay and Osgood argued that different types of disfluency represent different production problems or different strategies for correcting problems. But since this proposal, differences between disfluency types have received little examination. Experimental studies have often examined single types of disfluency without comparison across categories. Further exploration of the differences between disfluency types is necessary because many psycholinguistic studies have used disfluency to study language production (e.g. Levelt, 1983) or comprehension (e.g. Arnold et al., 2003; Ferreira and Bailey, 2004; Fox Tree, 1995). Without a generalized theory of the relationship between production and the various types of disfluency, it is not clear how well findings regarding a single type of disfluency generalize to others.

The present study investigates the differences between fillers, silent pauses, repairs, and repeats in an extended discourse. We argue that some of the differences in distribution between these disfluency types can be understood by considering them in the context of a model of the language production system. Most models of language production (for review, see Bock, 1995) posit at least three cascaded levels: a message level representing preverbal meaning, a grammatical level at which lexical items are selected and assembled into a morphosyntactic structure, and a phonological level at which an utterance's overall prosody and the phonological encodings of individual words are constructed. We provide evidence that disfluency types differ on at least two dimensions related to this system : (a) the temporal relation of the underlying problem to the current state of the production system, and (b) the level (or stage) of production at which the underlying problem occurred.

One dimension on which disfluency types may vary is when they occur relative to the underlying production difficulty that caused them. Some disfluencies may occur in response to a problem detected in already produced speech, while others may reflect problems in speech being planned. It is generally accepted that overt repairs are used when there is a problem in speech that was already produced (e.g., Levelt, 1983). Conversely, fillers have frequently been argued to reflect delays in planning or encoding of upcoming material (e.g. Arnold et al., 2003; Clark and Fox Tree, 2002), as have silent pauses (Butterworth, 1980). However, it has not been explicitly tested whether the distribution of these disfluencies differs from that of repairs.

The temporal properties of repeats are a matter of controversy. Clark and Wasow (1998) argue that repeats also reflect delays in planning and describe a *commit-and-repair* strategy: speakers commit to a partially planned utterance and, if planning delays prevent its initial fluent completion, they repeat the beginning so that the entire utterance can still be presented fluently. This theory predicts that the words repeated most often should be those likely to be produced during these early commitments, and Clark and Wasow find that function words, which tend to begin major constituents, are indeed repeated more often than content words. However, Levelt (1983) suggests that some repeats may actually result from "false alarms" of production monitoring systems. When the repair process is erroneously initiated in response to an acceptable utterance, the material in question ends up being reproduced without change, resulting in a repeat. This theory predicts that most repeats should share more properties with repairs. Since it is unknown whether either or both of these mechanisms underlie repeats, the present study examined the temporal properties of repeats as well.

In addition to their temporal relation to a production problem, disfluency types may also vary on a second dimension: the level of production at which the underlying problem occurs. If production involves a series of stages, as reviewed above, it is likely that errors and delays can occur at all the levels. Problems at different levels may give rise to disfluencies differing in form and time course.

Fillers may be particularly associated with delays in message-level planning because speakers may use them as deliberate linguistic signals of difficulty in their planning speech (Clark & Fox Tree, 2002). If fillers are indeed used deliberately, then they should require message-level planning. Such message-level revision should be easy when the difficulty arose on the message level, but may be more difficult when information about grammatical and phonological difficulties must first be sent back to the message level. Thus, fillers should be particularly apt to arise from delays in conceptualization rather than problems with grammatical or phonological planning. Evidence from the literature suggests that message-level planning can indeed play a role in filler production. Swerts (1998) observed that fillers occur more at stronger discourse boundaries than at weaker ones, because more planning is required to determine the next message. Similarly, speakers produce more fillers when answering questions about which they are less certain (Smith and Clark, 1993). These findings have been interpreted as indicating that planning demands associated with new or difficult topics result in a higher rate of fillers.

While fillers may be most apt to arise from message-level difficulties, silent pauses and repeats may arise from problems at all levels. Information about problems at the grammatical and phonological levels may not easily reach the message level to produce a revision or signal of difficulty for listeners. When the production system cannot easily produce any overt message-level signal of difficulty, then delays in production would instead be manifested as a silent pause or repeat. Again, evidence from the literature is inconclusive, but suggests that grammatical and phonological factors can play a role in

silent pause production. Maclay and Osgood (1959) observed that, while fillers usually occur between phrase boundaries, silent pauses tend to occur *within* phrases. Because the unit of message level planning has been argued to be at least an entire phrase (e.g., Garrett, 1988), phrase-internal disfluencies may reflect mostly delays in grammatical and phonological planning processes such as lexical or phonological retrieval. However, Reynolds and Paivio (1959) argue, based on effects of noun concreteness on silent pause production, that silent pauses may operate on the message planning level as well. These hypotheses have not been directly compared.

2 Present Work

Two hypotheses about disfluencies have been proposed: (a) disfluency types differ in their temporal relation to the underlying production difficulty, and (b) disfluency types differ in the level of production at which the difficulty occurred. The present work examines whether these hypothesized dimensions reflect actual differences in disfluency form and distribution during language production, and where specific disfluency types fall on these axes.

On the temporal relation dimension, it was hypothesized that fillers and silent pauses reflect problems with upcoming speech. Repairs were expected to reflect prior problems, by definition. Two competing hypotheses regarding repeats were also compared: the commit-and-repair theory predicts repeats to be more associated with upcoming problems, while the false-alarm theory predicts repeats to be more associated with prior problems. On the level-of-production dimension, it was hypothesized that fillers and appropriateness repairs usually reflect problems at the message level, while silent pauses and repairs correcting errors of form usually reflect problems at the grammatical or phonological level.

These questions were investigated in the context of an extended monologue to provide both a naturalistic situation and a discourse context for examining potential message-level effects on disfluency. Language may be produced quite differently in an extended monologue, yet disfluencies have rarely been examined experimentally in these situations. Conversely, disfluency use has sometimes been examined via corpus studies, but these observational studies lack the controls available in experimental work. Investigation of disfluencies in the context of an extended monologue provided a balance between naturalism and experimental control.

A storytelling paradigm was used in which participants were presented with stories and retold them to audiotape. The stories were three passages from *Alice's Adventures in Wonderland* (Carroll, 1865). Each passage was centered around a set of fourteen key points. Each key point was either a single action or two related actions crucial to the plot of the passage, such as *Alice finds a cake marked "EAT ME."*

3 Method

Ten University of Illinois undergraduates participated for course credit. All were native speakers of English between the ages of 18 and 22.

Participants read three passages, each approximately 2000 words, excerpted from *Alice's Adventures in Wonderland* (Carroll, 1865). Each passage was chosen to represent a distinct section of the plot that involved a number of discrete actions and that had a specific beginning and end. The three passages involved Alice getting trapped in a cave, Alice visiting the White Rabbit's house, and Alice meeting the Duchess.

Each participant read all three stories, presented in randomized order. For each passage, the participant first read the printed copy of the passage. Participants were told to read at their preferred speed and not to memorize all the events of the story since they would be receiving a list of key points to include. After reading the full story, the printed story was taken away and participants received the list of fourteen key points, printed in bullet-point format on a separate sheet. When participants indicated they were ready, the experimenter turned on a digital recorder and recorded the participant telling the story. Participants could consult the list of key points while speaking but were required to retell the story in their own words. Each recording continued until the speaker indicated to the experimenter that he or she was finished. The participant then repeated the process with the next story.

3.1 Transcription

The first author transcribed each retelling from the recordings. The transcripts were then scored for the beginning and end of each of four types of disfluency. A *filler* was any use of *uh*, *um*, *ah*, or *er*; in the uncommon case of several fillers in a row, each was coded as a separate instance. *Silent pauses* were the perception of a disfluent gap in the fluent speech stream, based on the

speaker's typical speech rate and the surrounding prosodic context. *Repeats* were one or more repetitions without modification of the same word, part of word, or string of words. *Repairs* were mid-utterance alterations of material already produced, including abandonment of the entire utterance (sometimes termed a *fresh start*; e.g., Bear et al., 1993). For reliability, a second observer also coded all the recordings for silent pauses. Only silent pauses coded by both observers were included in the final analysis.

Repairs were then subcategorized either as *error repairs*, which corrected identifiable lexical, phonetic, or syntactic errors, or as *appropriateness repairs*, which involved either a rewording of the same concept, the addition of a previously unstated fact, or the correction of a previous factual error. To assess reliability of these subcategorizations, a second observer, blind to the experimental hypotheses, scored all the repairs. Agreement between the two observers was good ($\kappa = .75$); where the observers disagreed, the first author's ratings were used.

Transcripts were also coded for the beginning and end of each key point. The beginning of a key point was coded at the first phrase introducing a fact from the printed bullet-point list. The key point continued through the last phrase regarding that bullet-point, at which point the end was coded. Typically, each key point was then followed by additional elaboration of the events or by explanations of how the key points related to each other in time within the story (e.g., "This went on for some time before..."). Such elaboration was not coded as part of the key point.

4 Initial Analyses

Participants were successful in retelling the stories, including a mean of 13.33 of the 14 key points ($SE = 0.84$) per passage.

Participants were also frequently disfluent. Collapsing across disfluency types, a mean ratio of 6.55 disfluencies per 100 words was observed. This ratio is close to past estimates of 6 disfluencies per 100 words (Fox Tree, 1995) suggesting that the present task yielded a typical sample of disfluencies. However, because the number of words spoken differed between participants, the raw frequency of disfluency is confounded with the total amount of speech. Consequently, we calculated the ratio of disfluency in proportion to the number of words.

5 Temporal Relation Analyses

5.1 Relation to Difficult Material

It was hypothesized that disfluency types would differ in their distribution relative to the demands of new key points. New key points were expected to be especially difficult at multiple levels of production: they introduce new plot elements to the story that may require additional discourse-level planning, and they often require access of new lexical, syntactic, and phonological forms. Thus, it was expected that fillers and silent pauses, hypothesized to reflect trouble with upcoming speech, should be more common before new key points than elsewhere. Since repairs reflect prior problems, however, they should be more common *after* new key points.

To test this hypothesis, transcripts were divided into four regions based on the key point codings (see section 3.1). The *Within* region comprised all words inside a key point. The *Before* region included the three words immediately before each new key point was introduced. The *After* region included the three words immediately after the end of each new key point. The *Between* region included all the words not inside or within three words of the introduction of a new key point. Mean rates of each type of disfluency in each region are presented in Table 1.

Type	Between	Within	Before	After
Filler	2.02	1.18	5.19	2.87
Silent Pause	2.47	1.81	5.18	4.15
Repair	1.24	1.16	1.26	2.48
Repeat	0.80	0.74	0.62	1.24

Table 1. Rate of disfluency per 100 words by location relative to key points.

A planned comparison indicated that, as predicted, fillers occurred at a higher rate in the Before region ($M = 5.19$ per 100 words) than in the three other regions ($M = 2.02$), $F_1(1,9) = 16.56$, $p < .001$, 97.5% CI of the difference.[1] $= \pm 1.85$. Silent pauses were also more prevalent before new key points ($M = 5.18$) than elsewhere ($M = 2.81$), $F_1(1,9) = 11.22$, $p < .01$, 97.5% CI of the difference $= \pm 1.68$.

[1] Because a second comparison (see section 6.1) was also conducted for the rate of each disfluency type, Bonferroni correction was applied to the 95% confidence intervals to avoid compounding the Type I error rate.

Repairs, as predicted, occurred at a significantly higher rate in the After region (M = 2.48 per 100 words) than in the three other regions (M = 1.22), $F_1(1,9) = 6.55$, $p < .05$, 97.5% CI of the difference = ± 1.18. Repairs were *not* more prevalent in the Before region (M = 1.24) than in the other regions (M = 1.63), $F_1(1,9) = 0.61$, p = .44, 97.5% CI of the difference = ± 1.11. These differences in distribution are consistent with the hypothesis that repairs reflect problems in prior speech but fillers and silent pauses reflect problems in speech being planned.

For repeats, no overall effect of region was observed, $F_1(3,9) = 0.80$, p = .51. Recall, however, that the commit-and-repair theory posits that repeats should be most common when a speaker has just begun production of problematic material. This theory predicts that repeats should be most common immediately after the start of key points, not immediately before. Consequently, an additional *Beginning* region was created, comprising the first three words after the beginning of each new key point. A planned comparison revealed that repeats occurred at a marginally significantly higher rate in the Beginning region (M = 1.29 per 100 words) than elsewhere (M = 0.77), $F_1(1,9) = 4.04$, p = .10, 97.5% CI of the difference = ± 0.62, supporting the commit-and-repair theory. These results suggest that while repeats, fillers, and silent pauses all relate to planning difficulties, they may differ in time course relative to this difficulty: fillers and silent pauses tend to reflect upcoming difficulty, while repairs may reflect more immediate difficulty.

5.2 Relation to Other Disfluencies

If fillers and silent pauses reflect upcoming problems, they should be more apt to occur *before* repairs, which reflect past problems, than *after* repairs. This pattern should hold even within a single utterance that does not cross key point boundaries, and provides an additional test of the temporal location dimension. To test this prediction, all utterances containing at least one repair were examined for other disfluencies within the same sentence. The rate of disfluency in the portion of the sentence before the *reparandum*—the problem being repaired—was compared to the rate of disfluency in the portion of the sentence after the conclusion of the repair. Rates of disfluency before and after repairs are presented in Table 2.

Type	Before Repairs	After Repairs
Filler	2.44	1.81
Silent Pause	3.75	2.40
Repeat	0.79	0.62

Table 2. Rate of disfluency before the first repair and after the last repair within a sentence.

Within a sentence, the rate of fillers was significantly greater before repairs than afterwards, $t_1(9) = 3.15$, $p < .05$, as predicted. Silent pauses were also more common before repairs than after them, $t_1(9) = 3.29$, $p < .01$. Repeats were only marginally more prevalent before repairs, $t_1(9) = 2.00$, p = .08.

6 Level of Production Analyses

6.1 Clauses and Disfluency

Fillers and silent pauses were both observed to occur more frequently before new key points. Because new key points almost always begin new clauses, it is possible that this distribution simply reflects the grammatical and phonological planning demands of new clauses. Alternatively, the message-level and discursive demands associated with a new story event may create an additional burden on the production system beyond the effect of a new clause.

These hypotheses can be tested by comparing the Before region with another region that contains clause boundaries but does not precede new key points. The After region was expected to also contain clause boundaries because the end of a key point generally represented the boundary between the introduction of a point and its elaboration. A paired samples t-test confirmed that, by participants, the prevalence of clause boundaries did not significantly differ between the beginning and end of key points, $t_1(9) = 1.868$, p = .10. By participants, 97.30% of Before (SE = 2.40%) regions contained a clause boundary and 95.07% of After regions (SE = 3.08%) contained a clause boundary.

Because both the Before and After regions contain clause boundaries, they were used to compare the introduction of new key points to a clause boundary baseline. A planned comparison by key points revealed that fillers were significantly more common in the Before region (M = 5.19 per 100 words) than in the After region (M = 2.87), 97.5% CI of the difference = ± 2.26. However, no significant difference in the rate of silent pauses was found between the Before re-

gion (M = 5.18) and the After region (M = 4.15), 97.5% CI of the difference = ± 2.05.

These data suggest that planning new topics or points within a discourse may be especially difficult and may lead to more fillers than planning other clauses. However, since silent pauses occurred equally frequently at all clause boundaries, they may be more associated with the grammatical planning that should occur at all clause boundaries.

6.2 Distribution within Utterances

Because language production is incremental, planning at the grammatical and phonological levels continues until nearly the end of an utterance (Butterworth, 1980). Semantic and discourse-level planning, on the other hand, is thought to represent an earlier level of production that is completed sooner than grammatical and phonological planning (e.g. Bock, 1995; Butterworth, 1980). Thus, disfluencies reflecting message-level difficulties should be more prevalent early in an utterance, before message-level planning has been completed, whereas grammatical and phonological difficulties should be found throughout an utterance.

To test this hypothesis, a simple division of early and late locations within an utterance was constructed by dividing each utterance in half according to the number of words. Table 3 presents mean rates of disfluency for the first half of an utterance versus the second half. (Repairs and repeats that spanned the midpoint of an utterance were excluded from this analysis.)

Type	First Half	Second Half
Filler	3.19	1.24
Silent Pause	3.76	1.50
Appropriateness Repair	0.52	0.21
Error Repair	0.34	0.24
Repeat	0.82	0.55

Table 3. Rate of disfluency per 100 words by location with an utterance.

As predicted, fillers were significantly more common in the first half of an utterance than in the second half, by participants, $t_1(9) = 3.48$, $p < .01$, as were appropriateness repairs, $t_1(9) = 3.18$, $p < .05$. This suggests that both fillers and appropriateness repairs are associated with the message level. Also as predicted, the rate of error repairs did not significantly differ between the first half and second half, $t_1(9) = 1.54$, $p = .16$, consistent with the hypothesis that these repairs reflect grammatical and phonological processes that continue throughout utterance production. The frequency of repeats also did not significantly differ between the first half and second half, $t_1(9) = 1.68$, $p = .13$. Since silent pauses were hypothesized to be most associated with the grammatical and phonological levels, they were also expected to be equally prevalent in both halves of the utterance. Contrary to this expectation, however, silent pauses were significantly more common in the first half of utterances, $t_1(9) = 4.52$, $p < .01$.

6.3 Correlation between Types

A speaker having difficulty at a particular level is likely to produce many disfluencies associated with that level over the course of the task. Thus, it was expected that a speaker's overall rate of use of a particular disfluency type would correlate with the rate of use of other disfluency types that stem from problems at the same level. Pearson's correlation coefficients correlated filler and silent pauses with both types of repairs. Participants' overall rate of filler use was significantly positively correlated with the rate of appropriateness repairs, $r = .70$, $p < .05$, Bonferroni corrected, but not with that of error repairs, $r = .44$, $p = .42$, supporting the hypothesis that fillers are associated with the message level. As expected, silent pauses were uncorrelated with the frequency of appropriateness repairs, $r = .13$, $p > .99$, Bonferroni corrected, but contrary to expectations, silent pauses were also uncorrelated with the frequency of error repairs, $r = .37$, $p = .60$.

6.4 Utterance Length

It was expected that longer utterances would place greater demands on grammatical and phonological planning but not on message planning. Butterworth (1980) has argued that lexical retrieval takes time and can easily result in disfluency. The number of content words (nouns, verbs, adjectives, and adverbs) may thus index an utterance's grammatical and phonological planning demands. (Function words such as prepositions and determiners have frequently been argued to be placed by fast syntactic processes unlikely to result in disfluency; e.g., Butterworth, 1980). Thus, it was expected that silent pauses and error repairs, both hypothesized to reflect grammatical and phonological difficulties, would occur at a greater rate in utterances with

more content words. However, utterance length does not necessarily index demands at the message level. The *Alice's Adventures in Wonderland* passages contain a number of semantically anomalous events such as a baby turning into a pig. Although conceptually difficult, these events can be expressed with lexically and syntactically simple utterances (e.g., "The baby turned into a pig.") that do not place great demands on grammatical and phonological planning. Since an utterance's length is likely to be less related to its difficulty at the message level, the rate of message-level disfluencies should be less related to utterance length.

A regression was conducted to determine whether greater grammatical and phonological planning demands, as reflected in utterance length, predicted rate of disfluency. Controlling for participants and number of function words, the number of content words was a significant predictor of the rate of silent pauses (standardized $\beta = .23$, $r = .11$, $t(576) = 2.75$, $p < .01$) and of error repairs ($\beta = .18$, $r = .09$, $t(576) = 2.09$, $p < .05$), but not of the rate of appropriateness repairs ($\beta = .10$, $r = .05$, $t(576) = 1.20$, $p = .23$), of fillers ($\beta = -.01$, $r < .01$, $t(576) = 0.14$, $p = .89$), or of repeats ($\beta = -.01$, $r < .01$, $t(576) = -0.10$, $p = .92$). These results support the hypothesis that silent pauses and error repairs tend to reflect difficulties at the grammatical and phonological level, whereas fillers and appropriateness repairs reflect difficulties at the message level and are less affected by increased grammatical and phonological planning demands.

7 Discussion

Observation of the distribution of disfluencies in a storytelling task supported both proposed dimensions. On the temporal relation dimension, fillers and silent pauses were found to be associated with problems in upcoming speech, repeats with more immediate upcoming difficulty, and repairs with past problems. On the level of production dimension, fillers and appropriateness repairs were found to be most associated with the message level, while error repairs were most associated with the grammatical and phonological level. Results generally suggest that silent pauses are associated with grammatical and phonological problems, though analysis of the position of silent pauses within an utterance also suggests possible message-level influences on silent pause production. Repeats were not found to be clearly associated with one level or another; it is possible that they can function at all levels. These results suggest that the distribution of disfluency types could be diagrammed in a model containing at least two dimensions, as in Figure 1.

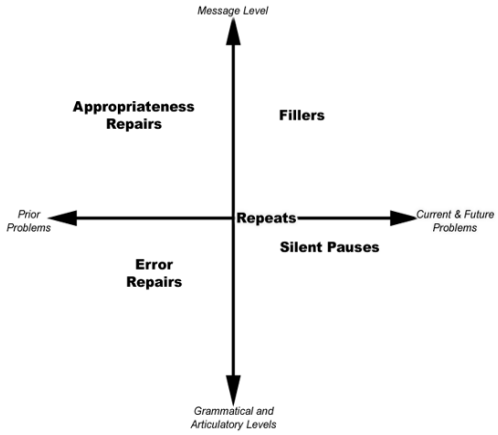


Figure 1. Schematic representation of the location of some disfluency types on dimensions of temporal location and level of production.

These findings suggest that current psycholinguistic work on language production provide a framework for understanding some aspects of disfluency. Of course, disfluency types may differ on other dimensions as well; in particular, some disfluencies may serve conversational purposes not captured by the present monologue task. For instance, fillers like *uh* and *um* have been argued to perform conversational functions like indicating a dispreferred response (e.g., Schegloff, 2006).

Nevertheless, the effects reported here of the message level of language production on disfluency are particularly important because message-level influences on disfluency are less frequently investigated. Prior findings that fillers occur more before objects with infrequent names (e.g. Schnadt and Corley, 2006) have often been interpreted as revealing an association between lexical access and filler production. However, uncommon objects differ from common ones in conceptual frequency as well as lexical frequency. Thus, prior studies confound the grammatical-level factor of lexical frequency with message-level semantic factors. The present study suggests that filler use may actually be more associated with message-level difficulty.

Why might fillers be most apt to arise from message-level difficulties? Recall that Clark and Fox Tree (2002) argue that speakers deliberately produce fillers to communicate the fact that they are having difficulty in production. Speakers

may desire to signal their difficulty for a number of reasons, such as self-presentation (Smith & Clark, 1993) or "holding the conversational floor" and preventing an interlocutor from beginning a turn (Maclay & Osgood, 1959). All these purposes should require message-level planning.

The distributions of disfluency observed in the present study also bear on the debate about what mechanisms underlie repeated words. Recall that Levelt (1983) has argued that repeats arise when the repair system is mistakenly activated and ends up reproducing the original utterance unmodified. Alternately, Clark and Wasow (1998) argue that repeats are part of a commit-and-repair strategy used by speakers to allow fluent delivery of an utterance that encountered planning problems early in delivery. In the present study, repeats occurred most frequently at the beginning of a key point. The commit-and-repair strategy predicts this pattern but the false alarm hypothesis makes no *a priori* prediction that repeats should follow such a distribution. Thus, the present data suggest the commit-and-repair theory may best describe most repeats.

7.1 Conclusion

Disfluency is not a unitary phenomenon. Speech in a discourse is subject to several types of disruptions, which can represent different problems and different responses from the production system. The present work demonstrates that fillers, silent pauses, repairs, and repeats differ on two dimensions related to language production: their temporal relations to the problem that caused them and in the level of production with which they are associated. These dimensions provide a framework by which the differences between various kinds of disfluencies can be captured in future psycholinguistic work on disfluency.

Acknowledgement

Scott H. Fraundorf was supported by National Science Foundation Graduate Research Fellowship 2007053221.

References

Jennifer E. Arnold, Maria Fagnano, and Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32(1): 25-36.

John Bear, John Dowding, Elizabeth Shriberg, and Patti Price. 1993. A system for labeling self-repairs in speech. SRI AI Center Tech Note #522.

Kathryn Bock. 1995. Sentence production: From mind to mouth. In Joanne L. Miller and Peter D. Eimas (Eds.), *Handbook of perception and cognition. Vol. 11: Speech, language, and communication* (pp. 181-216). Academic Press, Orlando, FL.

Brian Butterworth. 1980. Evidence from pauses in speech. In Brian Butterworth (Ed.), *Language Production, vol. 1: Speech and* talk (pp. 155-176). Academic Press: London, UK.

Lewis Carroll. 1865. *Alice's Adventures in Wonderland.* Retrieved September 15, 2006, from http://www.gutenberg.org/etext/11

Herbert H. Clark and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73-111.

Herbert H. Clark and Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology*, 37(3):201-242.

Fernanda Ferreira and Karl G. D. Bailey. 2004. Disfluencies and human language comprehension. *TRENDS in Cognitive Science*, 8(5):231-237.

Jean E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6): 709-738.

Merrill F. Garrett. 1988. Processes in language production. In Frederick J. Newmeyer (Ed.), *Linguistics: The Cambridge Survey: Vol 3. Language: Psychological and biological aspects* (pp 69-96). Cambridge University Press, Cambridge, UK.

Willem J. M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition,* 14(1):41-104.

Howard Maclay and Charles E. Osgood. 1959. Hesitation phenomena in spontaneous speech. *Word*, 14(1):19-44.

Allan Reynolds and Allan Paivio. 1968. Cognitive and emotional determinants of speech. *Canadian Journal of Psychology*, 22(3):164-175.

Emanuel A. Schegloff. 2006. On "uh" and "uhm" and some of the things they are used to do. Paper presented at BRANDIAL 2006, September 11-13, 2006, Potsdam, Germany.

Michael J. Schnadt and Martin Corley. 2006. The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the twenty-eighth meeting of the Cognitive Science Society.*

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32(1):25-38.

Marc Swerts. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4):485-496.

Timing in conversation: The anticipation of turn endings

Lilla Magyari
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
Lilla.Magyari@mpi.nl

Jan Peter de Ruiter
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
JanPeter.deRuiter@mpi.nl

Abstract

We examined how communicators can switch between speaker and listener role with such accurate timing. During conversations, the majority of role transitions happens with a gap or overlap of only a few hundred milliseconds. This suggests that listeners can predict when the turn of the current speaker is going to end. Our hypothesis is that listeners know *when* a turn ends because they know *how* it ends. Anticipating the last words of a turn can help the next speaker in predicting when the turn will end, and also in anticipating the content of the turn, so that an appropriate response can be prepared in advance. We used the stimuli material of an earlier experiment (De Ruiter, Mitterer & Enfield, 2006), in which subjects were listening to turns from natural conversations and had to press a button exactly when the turn they were listening to ended. In the present experiment, we investigated if the subjects can complete those turns when only an initial fragment of the turn is presented to them. We found that the subjects made better predictions about the last words of those turns that had more accurate responses in the earlier button press experiment.

1 Introduction

During conversations, a turn not only has to be relevant to the course of social interaction, but it also has to be appropriately timed. Sacks, Schegloff and Jefferson (1974) assume that transitions from one speaker to the next are accurately timed, so that gaps (silences between turns) and overlaps (i.e. when the interlocutors speak at the same time) are small. It is a normative rule of conversation that requires the participants to respond to the current speaker as soon as he/she has finished. When there are departures from this rule, the gaps or overlaps are interpreted communicatively. For example, a short silence before a response can indicate that the response is a disagreement when disagreement is a dispreferred action (Pomerantz, 1984).

Sacks et al.'s normative rule has been recently supported by measurement of floor transfer offset (FTO) in a data-set from Dutch two-party telephone conversations (De Ruiter et al., 2006). The FTO is defined as the difference between the time that a turn starts and the moment the previous turn ends. In the Dutch conversations, 45% of all speaker transitions had an FTO of between -250 and +250 ms, and 85% of them were between -750 and 750 ms. (Negative values indicate an overlap between the consecutive turns, positive values indicate a gap.) The FTO values were centered around 0. This pattern supports Sacks et al.'s (1974) assumption that gaps and overlaps are small. However, it also raises the question of how these accurately timed transitions are possible.

Such accurate temporal alignment of conversational turns suggests that a potential next speaker can anticipate the moment when the current turn is going to end. If the next speaker detects the end of the current turn (but she does not anticipates it), she will have a little delay before her turn because preparation for articulation requires some time. However, many turn transitions happen without temporal gaps.

Sacks et al. have already assumed that the potential next speakers can plan to align their turn accurately in time only if they are able to accurately predict the end of the current speakers turn. However, they left open the question of exactly how the anticipation of end of turns is carried out.

Many sources of information (semantic, syntactic, pragmatic and prosodic) have been proposed to be used in the prediction of turn endings. The few experimental studies which

have investigated this issue, mainly concentrated on the role of intonation in end-of-turn predictions. Groesjean and Hirt's study (1996) investigated if people can use prosodic information to predict end of French and English sentences. Subjects were listening to sentences that were presented in segments of increasing duration. They had to guess with how many words the fragments would continue. The sentences of which the initial fragment was presented to the subjects were either short or they were expanded by optional noun-phrases. The subjects had to guess using a multiple choice response task if the presented fragment was part of a short sentence or an expanded, longer sentence. The predictions did not improve with increasing duration of the fragments (sentence beginnings). Only when the first potentially last word was presented (i.e. the first point in the sentence where the sentence could end if it would be a short sentence) could the subjects predict if the sentence would be finished after the potentially last word or it would continue with 3 or 6 more words. According to Grosjean and Hirt the results indicate that in English prosodic information is made available for the prediction of sentence length only when the semantic and syntactic information can not help. Their similar experiment on French showed that subjects could tell if a sentence has ended or not. But they could not predict with how many words the sentences (3, 6 or 9 more words) would continue.

Grosjean and Hirt's study used recordings of sentences read aloud. The prosodic pattern may differ from the prosody occurring in natural conversations. Therefore, it is questionable how their results can be generalized to account for processing of spontaneous speech.

De Ruiter et al. (2006) investigated the contribution of the lexico-syntactic content and intonation in end-of-turn predictions. They manipulated recordings of natural conversations. Subjects listened to individual turns taken out from Dutch telephone conversations. They were asked to press a button exactly at the moment the turn ended. The duration between the end of the turn and the button-presses (called *bias*) was measured. In the different experimental conditions, the turns were presented naturally (as recorded) or a modified version was played. In one of the conditions, the intonational contour was removed, in another condition the lexico-syntactic content was removed by applying low-pass filtering. When subjects were listening to the original turns, their button-presses coincided with the turn-ends accurately; the distribution of the button-presses was similar to the distribution of FTO values for the same turns in the original conversations. There was no change in accuracy when the intonational contour was removed, but the performance got worse when the words could not be understood (note that the intonational information was still present in those stimuli). De Ruiter et al. concluded that the intonational contour is neither necessary nor sufficient for the prediction of turn-ends. These results suggest the lexico-syntactic information plays a major role in timing of turns.

Listeners have to perform many simultaneous tasks before they start their turn. They have to perceive and comprehend the current turn, and also formulate and time their subsequent utterance appropriately. The fine temporal alignment of conversational turns shows that these tasks have to be done simultaneously. Response preparation has to start before the previous turn ends in order to avoid gaps. Response preparation, however, can be initiated only if the speaker knows roughly what to respond. Therefore, the next speaker has to anticipate not only the end of the turns but also their content. When the last words of a turn can be anticipated they give information about the content and about the duration in advance. Therefore, we hypothesize that lexico-syntactic information helps in the prediction of the time when a turn will end through the anticipation of the last words of a turn. In other words: People know *when* a turn ends by knowing *how* it ends.

In order to test this hypothesis we conducted an experiment using the experimental stimuli of De Ruiter et al.'s study. Our prediction was that the more accurate the button-presses to the end of a given turn were in the earlier experiment, the more accurately the last words of that turn can be predicted. Therefore, we examined if there was any correlation between the accuracy of button presses in the earlier experiment and the off-line prediction of last words of the turn in a gating study. The end of selected turns were cut off at several points and fragments or the entire turn were presented to subjects who then had to guess how the turn would continue.

2 Methods

2.1 Participants

Fifty native speakers of Dutch (forty-two women and eight men, aged between eighteen and

twenty-nine) participated in the experiment. The data of one subject was excluded because the results showed that he did not understand the task correctly. The subjects were paid for their participation.

2.2 Stimulus material

The experimental materials were selected from stimuli used by De Ruiter et al. These stimuli were turns from natural conversations in Dutch. In the De Ruiter et al. experiment, it had been measured for each turn how accurately subjects could predict the end of turns by button-press. The temporal offset (bias) between the end of the turn and the button-presses was measured. The averaged bias of a turn indicates how accurately subjects could on average predict the time point of the end of that turn. A turn with a highly positive bias means that subjects pressed the button too late. A low bias (small positive value or with a small negative value) shows that subjects pressed the button on time or a bit earlier, just before the turn ended.

For the purposes of the present study, turns with high and low biases from the De Ruiter et al. study were selected. It was observed that turns with longer duration tend to have a lower bias. In order to avoid effects caused by the duration of the turns, ten turn-pairs were selected, where both members of the pairs had the same duration. The members of each pair were from different conversations produced by different speakers. The members of each pair had the same duration (max. difference between the members of the pairs was 16 ms), but they differed in their average bias. One of the members of every pair had higher average bias (between 237 and 123 ms), while the other member had a lower average bias (between -18 and 122 ms) relative to the other member of the pair. The durations of the 10 stimuli pairs were varying between 1.13 s and 2.05 s.

For each turn pair, four versions were made by cutting off the speech at four different temporal locations. The cut-off locations within each pair were at same points in time measured from the end of the recordings, but they were different across stimuli pairs. The cut-off locations were determined in a pair according to the boundaries of the two last words of each of the pairs. Each stimulus was cut at four points which were just at word boundaries at one of the members of a stimuli pair. The cut-off location varied across the pairs (the first points were on average at 0.76s from the end, the second points at 0.52s; the third points at 0.40 s; the fourth points at 0.25 s). Table 1. shows an example of gating points of one of the turn pairs that was used in the experiment. Turn A and B have almost the same duration (1.78 and 1.79 s), while A is a low bias turn (40 ms) and B is a high bias turn (226 ms). The vertical lines shows the points where both turns were cut in order to create the fragments. The vertical lines that are aligned with each other between the two turns indicate that the cut-off was made at the same points in time measured from the end of the recordings.

A.
maar dat hoor ik wel via |de|
mi|crof |oon
B.
ja maar daar moeten we maar een keer met| |
zijn|allen| heen

Table 1. Example of a turn-pair and the cut-off locations (shown by the vertical lines in the text)

2.3 Experimental design

Subjects were randomly assigned to one of five experimental lists. The stimuli in the lists were presented in random order to each subject. Their task was to type in if the presented segment constituted a complete turn. If the subjects decided that the turn was not complete, they were asked to guess and type in how they thought it would continue. If they did not have any guess about the continuation, they were asked to guess with how many words the turn would continue. They had to make a forced choice between A. one word, B. two words, or C. three or more words. Subjects were also asked how certain they were of their responses on a four point scale.

2.4 Procedure

The subjects were requested to sit in front of a computer screen and a keyboard with headphones. The instructions were visually presented on the screen. Before each stimulus a sentence was presented on the screen in Dutch, saying: "When you press the space bar you can listen to the next sound fragment two times.". 500 ms after pressing the space bar, a stimulus was presented two times, with a 1500 ms pause between the two presentations. After the stimulus presentation, the subjects saw a prompt (>:) on the screen where they had to type their guess

about the continuation of the fragment. If they thought the turn that they were listening to was complete, they had to type: '.'. If they did not have any guess about the continuation, but they did not think that the turn had finished, they were asked to type a '-'. After reading the instructions, the participants did a training session during which four stimuli were presented that were not part of the experimental list. After the training session, and possibly providing verbal clarifications, the experimenter left the room and the participants could continue the experiment alone.

2.5 Data-coding

Two variables with categories were created based on the responses. The variable PREDEND (prediction of the rest of the turn) was 0 when the continuation of the turn was entirely correct. It was 0 also if it was indicated correctly that the turn has ended. PREDEND could get 0 only if the guess was entirely correct regardless how many words had to be guessed. PREDEND was 1 when it was incorrect: when different words were used, when the end was indicated wrongly or when the participants did not have any guess.

PREDNUM (prediction of the number of words) variable had three categories: 1, when the predicted number of words was the same compared to the original version of the sentence even if the words were not the same, or when the participant did not have any idea about the continuation (but the prediction of the number of words in the continuation was correct); 2, when the predicted number of words was less than the number of words in the turn, and 3, when more words were predicted.

Responses which were not clear (e.g. words that do not exist) were excluded from the analysis. Only 3% of the data points were excluded.

3 Results

3. 1 Statistical analysis

The results were analyzed using a generalized linear mixed effects model (GLMM) (Baayen, 2008, Pinherio & Bates, 2000). We used this statistical analysis because of two main reasons. On one hand, it has been shown that mixed-effects models provide a better method for statistical analysis with repeated measurement data (see for example, Baayen, Davidson & Bates, 2008). Among other advantages, the mixed effect regression model can simultaneously handle all factors that potentially can contribute to explaining the variance in the data. The model can include random effects, such as variations caused by individual differences among the subjects and variations caused by differences in the properties of the items. It is also possible to fit the model to unbalanced data.

GLMM also has many advantages over the widely used repeated-measurement analysis of variance (ANOVA) for categorical datasets. In this experiment, the proportions of correct and non-correct responses were analyzed that do not follow normal distribution that can be problematic for the ANOVA analysis. When ANOVA is used for categorical outcomes, it can yield spurious results that GLMM can avoid (Jaeger, 2008).

3.2 Recognition of turn-ends

Figure 1 shows the percentage of correct responses when subjects were listening to the entire turn. The responses are highly accurate. 96% of the participants give correct responses at high-bias turns, and 90% of the participants at low bias turns.

The PREDEND variable was binary (correct or not correct), therefore a binomial distribution was specified for the model.

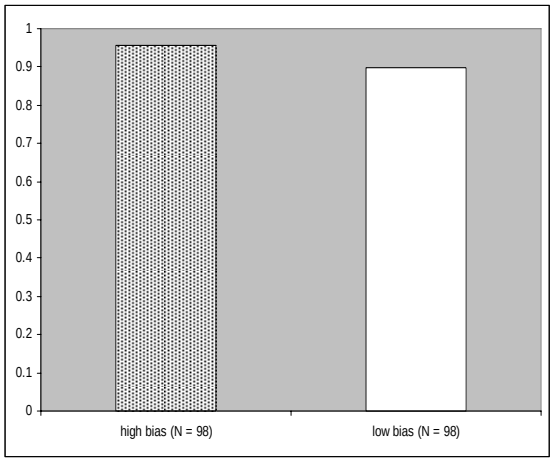


Figure 1. The proportion of correct responses ('the turn has ended') when the entire turn was presented

The linear model had Bias (if the turn belonged to the high or low bias turns) as a fixed effect, and Subjects and Utterance-pairs as

random effects. The GLMM analysis did not show any effect of Bias ($z = 1.498$, $p>0.1$, $N = 196$).

3.3 Prediction of the continuations

Figure 2 shows the proportion of the correct continuations at each cut-off location for both turn types. From the first cut-off location (I) to the fourth (IV) increasing proportion of the turns were presented to the subjects. The proportion of correct answers is increasing as the presented fragments get longer.

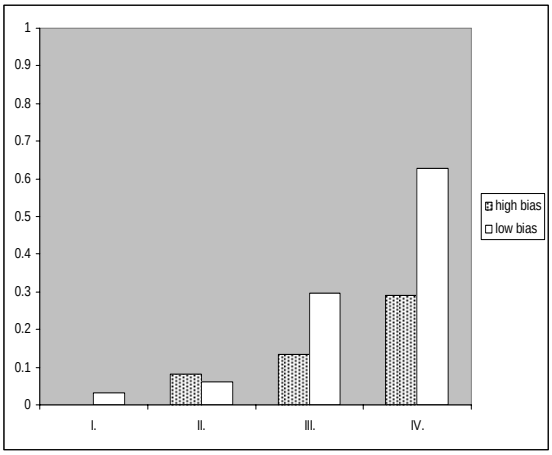


Figure 2. Proportion of the correct continuations at each cut-off locations. The x-axis shows the proportion (between 0 and 1), the y-axis shows the cut-off locations (from I. to IV. the duration of fragments from each turn are increasing). The white columns show the proportion of correct continuations when a fragment from a low bias turn was presented, the grey columns show the proportion of correct answers that belong to the high bias turns.

However, it is possible that differences between the two turn types may arise from the properties of the stimuli material. Some fragments were cut so close to the end of the last word that it sounded as the end. Therefore, the correct response was that the turn has ended and not a free guess about the continuation. It is probably easier to decide if a turn continues or not than it is to predict its continuation. Therefore, those turns where despite of the cut off, there was no more reliable auditory information coming, were excluded from our analysis (11%). Figure 3 shows the proportion of the correct responses at the four consecutive cut-off locations after the exclusion.

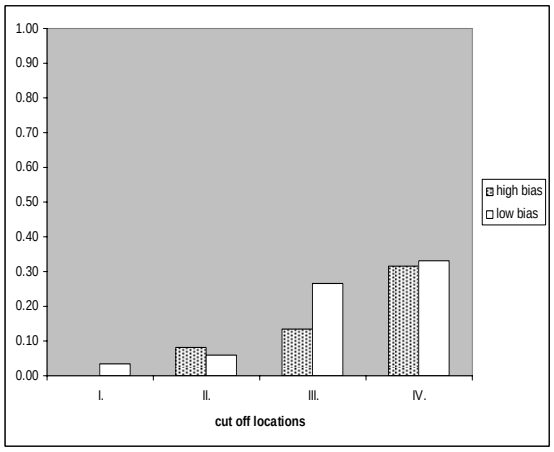


Figure 3. Proportion of correct continuations after excluding some of the fragments. For explanation of the figure, see Figure 2.

The differences between the turn types got reduced at the last cut-off location (IV). Table 2 shows the number of items at each cut-off location for both turn types and the proportion of the correct responses.

bias	I.	II.	III.	IV.
high	0.00 (N=96)	0.08 (N=98)	0.13 (N=98)	0.32 (N=68)
low	0.03 (N=98)	0.06 (N=98)	0.27 (N=78)	0.33 (N=60)

Table 2. The proportion of the correct continuations (1 = all are correct) and the number of items at each cut-off locations for both turn-types

At the last two cut off locations (III and IV) at 20% and at 31.5% of the cases subject were able to guess the correct continuations.

The GLMM had Bias and Cut-off location as fixed effects, and Subjects and Utterance-pairs as random effects. Table 3. shows the β-coefficients of the fixed effects in the model.

Predictor	Coeff	SE	z value	p
Intercept	6.578	0.774	8.503	<0.001
Cutoff	-1.318	0.166	-7.962	<0.001
Bias	-0.674	0.302	-2.235	<0.05

Table 3. The summary of the fixed effects in the GLMM of correct continuations at the four cut-off locations. (Coeff = Coefficient)

Both the cut-off locations (z -7.962, $p<0.001$, $N=694$) and the bias ($z=-2.235$, $p<0.05$, $N=694$) had a significant effect on the correct responses. It is possible, however, that low bias turns were

easier to complete because they always ended in a longer word than high bias turns. It means for example, that at the last cut-off location (IV) the fragments ended during the last word at the low bias turns, while the fragments ended before the last word at the high bias turns. In this case, maybe it is easier to recognize a word that was partially played than to guess for a not-heard at all word. In order to explore if this explanation is valid, the proportion of fragments that ended during the last word and before the last word at the fourth cut-off location was calculated for both types of turns (Figure 4).

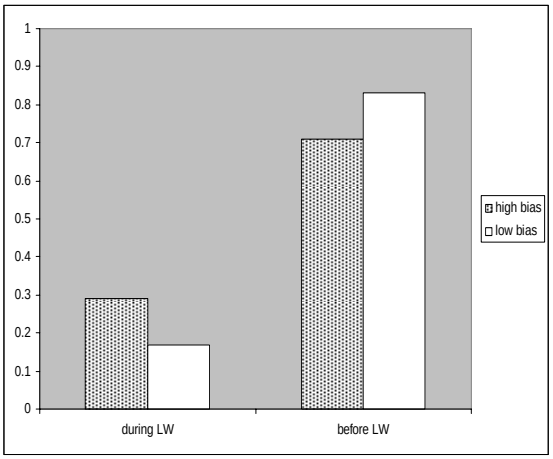


Figure 4. Proportion of stimuli that was cut during the last word (LW) or before the last word at the last cut-off location

The cut-off locations occurred during the last word in a smaller percentage at the low bias turns than at the high bias turns. The GLMM shows a significant effect ($z=2.5375$, $p<0.05$, $N=128$) of the position of the cut-off location (WB, during or before the last word) on the correct continuations. However, the direction of the effect is in the opposite direction. The guesses get better when the cut-off location is before the last word and not during it. Therefore, the observed differences between the turn types can not have been caused by the earlier recognition of the last words.

3.4 Prediction of the number of words

Our question was also if the difference between the high and low bias turns is not only caused by anticipation of the correct turn endings but also by the prediction of the correct number of words, irrespective of their form or meaning. We examined if there is a correlation between the turn types and the expectations about the length of the turn even if the continuations are wrong. Therefore, the number of the predicted words (PREDNUM) was analyzed for the cases where the prediction of the actual continuation was not correct. We again excluded those turns that has already finished at the two last cut-off locations.

Figure 5 shows the proportion of correct guesses about the number of words. We expected to find a higher proportion of correct responses at the low bias turns because that could lead to accurate button-presses. But Figure 5 shows the opposite direction of the differences.

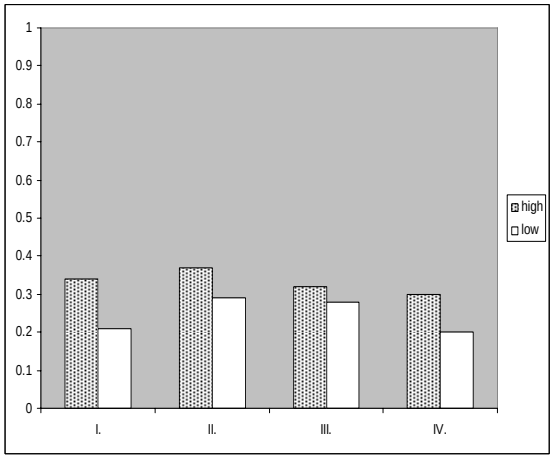


Figure 5. Proportion of correct guesses of the number of the coming words among the wrong continuations. For more explanation of the figure, see Figure 2.

Bias and the Cut-off locations were included as fixed effects, while Subjects and Utterance-pairs were included as random effects in the linear mixed effects regression analysis of the correct number of words estimates among the wrong guesses. The analysis showed a main effect of Bias ($z=2.56$, $p<0.05$, $N=601$) (Table 4).

Predictor	Coefficient	SE	z value	p
Intercept	0.611	0.265	2.307	$p<0.05$
Bias	0.476	0.186	2.56	$p<0.05$
Cutoff	0.035	0.088	0.399	$p>0.05$

Table 4. The summary of the fixed effects in the GLMM of correct estimates of the number of words among the wrong guesses

Figure 6 shows the proportion of the responses that predicted less number of words than the number of words that were still coming but not played.

The GLMM analysis (Table 5) showed that there is a significant effect of the Cut-off locations ($z=5.029$, $p<0.001$, $N=601$), and that

there is an interaction between Bias and Cut-off locations (z=-5.071, p<0.001, N=601). The difference between turn types in the less number of words predictions are increasing towards the end of the turn. The low bias turns tend to have a higher proportion of less number of words guesses.

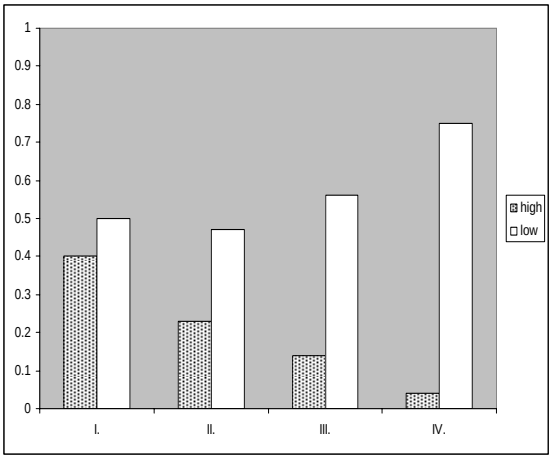


Figure 6. Proportion of guesses predicting less number of coming words among the wrong continuations. For more explanation of the figure, see Figure 2.

Predictor	Coeff	SE	z value	p
Intercept	-0.394	0.402	-0.98	p>0.05
Bias	0.588	0.445	1.322	p>0.05
Cutoff	0.828	0.165	5.029	p<0.001
Interaction: Bias&Cutoff	-1.06	0.209	-5.071	p<0.001

Table 5. The summary of the fixed effects in the GLMM of estimates of less number of words among the wrong guesses. (Coeff = Coefficient)

Figure 7 shows the proportion of the responses that predicted more number of words than the number of words that were still coming but not presented. The regression analysis showed an effect of Bias (z=-2.154, p<0.05, N=601) and Cut-off locations (z=-4.895, p<0.001, N=601), and also their interaction (z=4.836, p<0.001, N=601). More number of words were predicted at the high bias turns than at the low bias turns (Table 6).

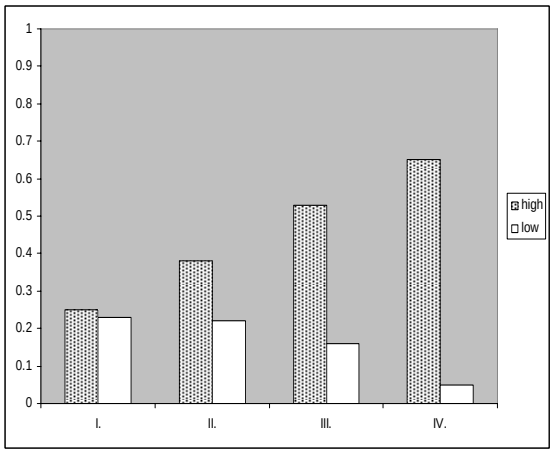


Figure 7. Proportion of guesses predicting more number of coming words among the wrong continuations. For more explanation of the figure, see Figure 2.

Predictor	Coeff	SE	z value	p
Intercept	1.655	0.316	5.242	p<0.001
Bias	-0.987	0.458	-2.154	p<0.05
Cutoff	-0.582	0.119	-4.895	p<0.001
Interaction: Bias&Cutoff	0.983	0.203	4.836	p<0.001

Table 6. The summary of the fixed effects in the GLMM of estimates of more number of words among the wrong guesses. (Coeff = Coefficient)

In order to see how early the differences in the number of words predictions are present, an additional analysis was done for the first two cut-off locations (I and II). Fragments with these cut-off locations ended before the last word in all cases. A mixed effect regression model was fitted for the cut-off location I and II separately with Bias as main effect, and Utterance pairs and Subjects as random effects. Bias had an effect at both cut-off locations when less number of words were predicted: At location I, z=-2.187, p<0.05, N=191 and at location II, z=-3.647, p<0.01, N=182. Bias did not have an effect at the first cut-off location (z=0.298, p>0.05, N=191), but it had an effect at the second cut-off location (z=2.35, p<0.05, N=182) when more number of words were predicted. This means that the subject predicted in a higher proportion less number of words at the low bias turns, and more number of words at the high bias turns by listening to fragments that did not contain the last word of the turns.

4 Discussion

We investigated the hypothesis that people know when a turn ends because they know how it ends. Gating paradigm was used to examine if it is possible to predict the last words of conversational turns. The turns were extracted from natural conversations and the original context was not presented to the subjects. Even so the subjects could guess the not presented or only partially presented last words of the turns correctly in around 20 - 30% of the cases. We found differences also among the turn-types. The continuation of turns whose ends were indicated too late by the button-pressing task in an earlier experiment were less often predicted correctly than the continuations of those turns whose ends were indicated on time (low bias turns). We have shown that these differences between turn-types could not have been caused by earlier or later recognition of the last word of that turn.

We also found that when the continuations were not correct subjects predicted less numbers of words at the low bias turns, and more numbers of words at the high bias turns before the last word of a turn. This shows that probably when the subjects thought that more words were coming, they pressed the button too late in the button-press experiment. These results support the hypothesis that the prediction of turn endings is based on the predictions made about the content and word forms of the turn.

This study emphasis the role of anticipation in order to explain the alignment of turns during conversations. This is in line with studies that show that people use the linguistic context for anticipating the upcoming words (DeLong, Urbach & Kutas, 2005). Pickering and Garrod (2007) argue that comprehenders use the production system for making predictions in order to achieve a faster and easier comprehension. However, investigation of conversational turn alignments shows that it is very likely that preparation of responses occurs in temporal overlap with the comprehension or prediction of the current turn. If the production system is used for predicting the upcoming words, then the same system is used also for preparation of the coming turn. We doubt whether the same system can fulfill two tasks at the same time. We think it is more plausible to assume that the comprehension system is used for anticipation of the content and word forms of the current turn, while the production system is used for preparation of the next turn.

This off-line study has some limitations in generalizing its results to on-line processing. However, the study shows that people can make accurate predictions about the final word forms of turns from natural conversations. The results also suggest that anticipation about the number of words of a turn can explain the accurate performance in turn-end predictions.

References

Katherine A. DeLong, Thomas P. Urbach and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8:1117-1121.

Harald Baayen. 2008. *Analyzing Linguistic data: A practical introduction to statistics.* Cambridge University Press.

Harald Baayen, Doug Davidson and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, in press, doi:10.1016/j.jml.2007.12.005.

Francois Grosjean and Cendrine Hirt. 1996. Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and Cognitive Processes*, 11:107-34.

T. Florian Jaeger. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, in press, doi:10.1016/j.jml.2007.11.007.

Martin J. Pickering and Simon Garrod. 2007. Do people use language production to make predictions during comprehension?. *TRENDS in Cognitive Sciences*, 11:105-110.

J. C. Pinherio and D. M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer.

Anita Pomerantz. 1984. Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes. In J. M. Atkinson and J. Heritage (Eds.). *Structures of Social Action*. Cambridge, Cambridge University Press, 53-101.

J. P. de Ruiter, Holger Mitterer, and Nick J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation, *Language*, 82:515-535.

Harvey Sacks, Emanuel A. Schegloff and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696-735.

Part II

Posters

Adaptive Natural Language Generation in Dialogue using Reinforcement Learning

Oliver Lemon
School of Informatics
Edinburgh University
olemon@inf.ed.ac.uk
http://homepages.inf.ed.ac.uk/olemon

Abstract

This paper presents a new model for adaptive Natural Language Generation (NLG) in dialogue, showing how NLG problems can be approached as statistical planning problems using Reinforcement Learning. This approach brings a number of theoretical and practical benefits such as fine-grained adaptation, generalization, and automatic (global) optimization. We present the model and related work in statistical/trainable NLG, discuss its applications, and provide a demonstration of the approach, showing policy learning for adaptive information presentation decisions (Contrast, Cluster, or List items). An adaptive NLG policy learned in our framework shows a statistically significant 27% relative increase in reward over an "RL-majority" baseline policy for the same task. We thereby also show that that such NLG problems should be approached in combination with dialogue management decisions, and we show how to jointly optimize NLG and dialogue management plans.

1 Introduction

Natural Language Generation (NLG) in dialogue is often characterised as choosing "how" to say something once "what to say" has been determined. In principle, NLG in dialogue thus comprises a wide variety of decisions, ranging over content structuring, choice of referring expressions, use of ellipsis, aggregation, and choice of syntactic structure, to choice of intonation markers. In computational dialogue systems, "what to say" is usually determined by a dialogue manager (DM) component, via planning, hand-coded rules, finite state machines, or learned policies, and "how to say it" is then very often defined by simple templates or hand-coded rules which define appropriate word strings to be sent to a speech synthesizer or screen.

Previous statistical approaches to NLG are reviewed in section 2, but none of them have explored NLG as statistical *planning*. Some aspects of NLG have been treated as planning (Koller and Stone, 2007; Stone et al., 2003), but not statistically. Prior dialogue-related work in statistical planning (e.g. Reinforcement Learning) has dealt only with policies for planning dialogue acts in an information gathering phase, and has not been applied to NLG decisions themselves (though see Rieser and Lemon (2008) for work in multimodal generation).

Learning approaches have several key potential advantages over template-based and rule-based approaches to NLG (we discuss "trainable" NLG in section 2.1) in dialogue systems:

- ability to adapt to fine-grained changes in dialogue context
- data-driven development cycle, with reduced development costs for industry.
- provably optimal action policies with a precise mathematical model for action selection
- ability to generalize to unseen dialogue states

We aim to illustrate these advantages in the demonstration system described in this paper.

2 Prior work

There are 3 main approaches to generating system utterances in dialogue systems: template-based NLG, conventional NLG as developed in the text generation literature (Reiter and Dale, 2000), and more recently, trainable generation (Barzilay and Lapata, 2005; Duboue and McKeown, 2003; Stent et al., 2004; Walker et al., 2001; Walker et al., 2007).

Template-based generation is typically used in industrial dialogue systems, and even in most state-of-the-art research systems. However, this approach requires that new templates be created by hand for each application, and severely limits that system's ability to adapt to dialogue context or user preferences, due to the practical constraints of having to write different templates for each possible combination of feature values.

Conventional NLG typically follows a pipeline architecture consisting of three main modules (Reiter and Dale, 2000):

- (i) a text planner which performs content selection and discourse structuring,
- (ii) a sentence planner which selects attributes for referring expressions, aggregates content into sentence-size units, and selects lexical items, and
- (iii) a surface realizer which converts sentence plans into natural language.

This approach has been successfully applied in systems that tailor their presentations to the user's preferences (Carenini and Moore, 2006; Demberg and Moore, 2006; Moore et al., 2004; Walker et al., 2004) and to the dialogue context (Isard et al., 2003), but these systems generally use rules that are specifically hand-crafted for a particular domain. A major problem with these standard NLG approaches is that hand-coded rules, manually-set thresholds, and templates all severely limit the adaptivity that can be achieved in NLG, both in the amount of adaptivity possible and the ability to adapt to fine-grained changes in the dialogue context, user behaviour, or environment (e.g. noise levels). The standard approaches are limited by the expertise of the system designer, and the adaptivity that they can encode in their rules or templates. Statistical approaches in general promise a more practical, effective, and theoretically well-founded approach to adaptivity in NLG, because they are not limited by human design capacities, and can be trained from data. Conventional NLG approaches can also be too slow for real-time dialogue applications (Stent et al., 2004). There has therefore been recent interest in statistical methods in the area of "trainable" NLG.

2.1 Trainable NLG

Trainable NLG is a more recent approach, where automatic techniques are used to train NLG modules, or to adapt them to specific domains and/or types of user. However, this work has focussed on local optimization through supervised learning, and has not explored global decision-theoretic planning approaches such as Reinforcement Learning.

Early work here focused on supervised learning of how to produce surface forms from sentence plans, using overgeneration and ranking, using either bigram language models (Oh and Rudnicky, 2002), or ranking rules learned from a corpus of manually ranked training examples (Walker et al., 2001). More recent work has extended this approach to sentence-planning (Stent et al., 2004).

In this work, given a content plan (the propositions to express and discourse relations among them), a generator first produces a set of text-plan trees, consisting of speech acts to be communicated, and the rhetorical relations between them. For each of these, a set of candidate sentence plans are generated by a heuristically ordered set of clause-combining operations. The sentence plan ranker is then trained by using the RankBoost algorithm to learn a set of ranking rules from a manually labelled set of examples.

For content selection, recent research has shown that given a corpus of texts and the database of facts or events it describes, content selection rules can be learned (Barzilay and Lapata, 2005; Duboue and McKeown, 2003). In this work, content selection has been treated as a binary classification task. Here, semantic units in the database are first aligned with sentences in the corpus, and then classification is used to learn whether or not a semantic unit should be included in the text.

However, as explained above, these types of supervised learning used for NLG do not model the

required optimization and planning of sequences of actions-in-context which we propose to capture with RL techniques. An interesting issue for future work is how these types of classifier-based learning can be integrated with the MDP approach propsosed here.

3 The model: NLG as statistical planning

This paper treats NLG as a statistical planning and optimization problem using decision theory in the framework of Markov Decision Processes (MDPs), similar to (Rieser and Lemon, 2008). The main advance here is to treat aspects of NLG within the same MDP-based planning and learning frameworks as have been successfully applied in speech recognition and dialogue management, for example (Levin and Pieraccini, 1997; Walker et al., 1998; Singh et al., 2002; Young, 2000).

We now propose to model the NLG problem as a Markov Decision Process (MDP). Here a stochastic system interacting with its environment (in our case, the user of the dialogue system) through its actions is described by a number of states $\{s_i\}$ in which a given number of actions $\{a_j\}$ can be performed. In a dialogue system, the states represent the possible dialogue contexts (e.g. how much information we have so far obtained from the user) and the actions are now system dialogue and NLG actions.

Each state-action pair is associated with a transition probability $\mathcal{T}_{ss'}$: the probability of moving from state s at time t to state s' at time $t+1$ after having performed action a when in state s. This transition is also associated with a reinforcement signal (or reward) r_{t+1} describing how good the result of action a was when performed in state s. In dialogue these reward signals are most often associated with task completion and dialogue length, but we will also associate them with NLG decisions.

To control a system described in this way, one then needs a strategy or policy π mapping all states to actions: $\pi(s) = P(a|s)$ In this framework, a Reinforcement Learning agent is a system aiming at optimally mapping states to actions, i.e. finding the best strategy so as to maximize an overall reward R which is a function (most often a weighted sum) of all the immediate rewards. In dialogue (and many other problems) the reward for an action is often not immediate, but is *delayed* until successful completion of a task. In the most challenging cases, actions may affect not only immediate reward, but also the next situation and, via that, all subsequent rewards.

In general, then, we are trying to find an action policy π which maximises the value $Q^\pi(s,a)$ of choosing action a in state s, which is given by the Bellman equation:

$$Q^\pi(s,a) = \sum_{s'} \mathcal{T}_{ss'}[\mathcal{R}_{ss'} + \gamma V^\pi(s')] \qquad (1)$$

(Here we denote the expected immediate reward by $\mathcal{R}_{ss'}$, γ is a discount factor between 0 and 1, $V^\pi(s)$ is the value of state s' according to π, see (Sutton and Barto, 1998)).

If the transition probabilities are known, an analytic solution can be computed by dynamic programming. Otherwise the system has to learn the optimal strategy by a trial-and-error process, for example using Reinforcement Learning methods (Sutton and Barto, 1998) as we do in this paper. Trial-and-error search and delayed rewards are the two main features of Reinforcement Learning.

In prior work on dialogue strategy learning, only dialogue acts (e.g. greet, ask slot, explicit confirm) chosen by the system have been optimized. Here we go beneath the level of dialogue acts to plan NLG actions such as content structuring. Several key questions thus arise – how to represent different NLG actions for planning, what context features and state representations are important in the MDP, and what reward signals can be used to optimize NLG?

We now present a fully worked example to show the model in use.

4 Learning for Adaptive Information Presentation

One of the classic problems in NLG is how to present one or more items to a user, for example by simply listing them, contrasting them in respect of some attributes, or clustering similar items together (e.g. (Moore et al., 2004; Walker et al., 2007)). For example the system may present items like so:

- LIST: "There are four hotels meeting your criteria. The first is the Royal, the second is ..."

- CONTRAST: “The Oak is an expensive central hotel. The Royal is cheap but is not central. . . .”
- CLUSTER: “There are 7 expensive hotels and 11 cheap ones, . . .”.

We will model these decisions in an MDP, and solve it using trial-and-error exploration, using Reinforcement Learning methods. First, we model the states of the system.

4.1 State space

In this example we will have 3 search constraint slots that the user can fill (for instance `food type, location, price range` for a restaurant search application, or `artist, album, genre` for music browsing). These slots can be either filled or confirmed. Confirmed slots have 100% chance of being correct, and filled slots only 80% chance, thereby modeling noise in the speech recognition environment (see also Lemon and Liu (2007), Rieser and Lemon (2008)). In addition we will model the number of “hits” or search results returned by the system after every user turn – this will be a number from 0 to 100.

4.2 Action set

See figure 1 for the hierarchical structure of the actions available to the system, representing the combined dialogue management (DM) and NLG task as an inter-related planning problem.

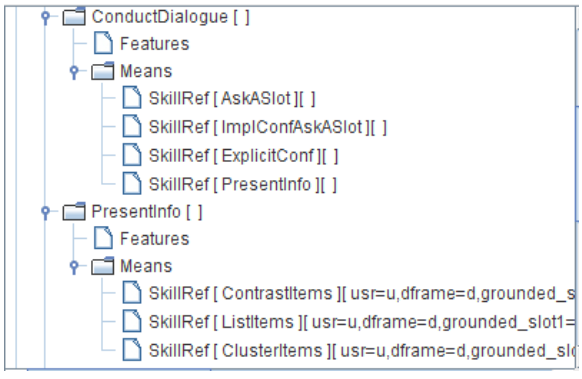


Figure 1: A Hierarchical Plan for NLG and DM

Here we see a top-level “skill” (ConductDialogue) responsible for dialogue management choices in the MDP, and a second level skill (PresentInfo) which is reponsible for the NLG choices. ConductDialogue governs the standard dialogue management options, and decomposes into the 4 possible action choices (or “Means”) AskASlot, ImplicitConfirm_and_AskASlot, ExplicitConfirm, and PresentInfo. This allows the system to choose between these 4 types of dialogue act at any time. More interestingly, for the NLG component of the system we have implemented possible 3 action choices (or “Means”) for information presentation under PresentInfo: ContrastItems, ListItems, and ClusterItems. ListItems is just the standard list content structuring operator, while ContrastItems and ClusterItems are actions which structure the items presented to the users by contrasting them and clustering them respectively, as shown above.

4.3 Reward function

Now that we have our states and actions, we need to define a Reward signal (or “Objective function”) for the learning system. This directs the learner in terms of its overall goals (e.g. short dialogues where users rate information presentation highly), while it is up to the learner to find an action policy which meets these goals. What makes the learning problem interesting is that these goals contain conflicting “trade-offs” that the system must learn to balance, based on the state that it is in. For example, the goal to have short dialogues conflicts with the goal to get reliable (i.e. confirmed) search constraints from the user and to present small numbers of items. The learning problem here is then for the system to decide at each turn whether to ask for more information/constraints, confirm (explicitly or implicitly) the exisiting information/constraints, or to List, Contrast, or Cluster the current items returned from the DB. Note that the system can decide to immediately present information in some way to the user even if not all slots are filled or confirmed. This leaves open the option for the system to exploit a “good” information presentation situation (such as having only 2 database items to tell the user about via a Contrast) even if the DM situation (e.g. having only 2 filled slots) is not in itself very rewarding. In this way the NLG and DM decisions are jointly optimised in this setup.

For training a system to be deployed with real

users, this reward/objective function would be developed based on a PARADISE-style (Walker et al., 2000) analysis of a small amount of Wizard-of-Oz data (Walker et al., 1998; Rieser and Lemon, 2008). Here, to prove the concept of statistical planning for NLG, we simply show that the learner can jointly optimize the NLG and dialogue management decisions based on a complex reward signal. Nothing depends on the particular values chosen here – they are for illustration only and can be estimated from suitable data.

The overall reward for each dialogue conducted by the system has 3 components: *completion reward*, *turn penalty*, and *presentation reward/penalty*. Turn penalty is simply -1 per system turn. The completion reward is the % probability that the items presented to the user correctly meet their actual search constraints, and is therefore a function of the number of filled or confirmed slots. For example, if all 3 slots are confirmed, then (in this noise model) we have 100% chance of having the search constraints correct. The number of filled/confirmed slots stochastically determines the number of items that the system can present to the user if it decides to enter the presentation phase. For example if 3 slots are filled, then 0-10 items will be presented to the user, if 2 slots are filled 0-20 items are retrieved, if 1 slot, 0-100 items. Again, in a real application, these distributions would be estimated from data.

The presentation reward (PR) for each information presentation action is defined as follows, for $i =$ number of items to be presented:

- ListItems: $0 \leq i \leq 3 : PR = 100; 4 \leq i \leq 8 : PR = 0; 9 \leq i : PR = -100$
- ContrastItems: $i \leq 1 : PR = -100; 2 \leq i \leq 6 : PR = 300; 7 \leq i : PR = -100$
- ClusterItems: $0 \leq i \leq 5 : PR = -100; 5 \leq i \leq 8 : PR = 0; 9 \leq i : PR = 300$

This range of rewards/penalties, together with those for filled and confirmed slots and dialogue length provides a complex environment within which the learner must explore different trade-offs.

4.4 Environment and User simulation

For training a policy given this definition of states, actions, and rewards, we also need an environment simulation that responds appropriately to system actions. Here the environment is not only the user, but also the database from which items are retrieved for presentation to the user. For policy exploration we use a simple bigram stochastic user simulation with probabilities estimated from COMMUNICATOR data, similar to (Georgila et al., 2006). At each system turn, a number of database hits is randomly determined as a function of the number of filled search constraint slots, as described above. Note that this user simulation does not need to respond directly to the NLG decisions of the system, since the dialogue closes as soon as the system decides to present information (in whatever manner) to the user. A central open question for this type of MDP model of NLG is how to develop "good" user simulations that are sensitive to system NLG choices (Janarthanam and Lemon, 2008).

4.5 The "RL-majority" Baseline policy

In contrast to other work on policy learning, which only uses hand-coded systems for comparison, we choose a more challenging baseline. This is because hand-coded policies have been shown to be inferior to learned policies in numerous studies, e.g. (Levin and Pieraccini, 1997; Singh et al., 2002; Lemon and Liu, 2007; Walker et al., 1998), and also, because our task here is a combination of dialogue management and NLG, we do not want the NLG results to be contaminated by an inferior hand-coded dialogue management policy. We therefore choose to compare against a baseline policy learned for the same problem domain, but where the learner uses the *average* most rewarding action for the NLG component (in this case, Cluster items). We call this the "RL-majority" baseline, because it is the RL analogue of a majority class baseline. This baseline policy does not have access to the "DB hits" feature for decisions under PresentInfo (it does have this feature for the top level decisions though), so it learns the average best NLG action rather than attempting to learn the best NLG action for each possible number of DB hits.

4.6 Training the policies

We use a hierarchical SARSA Reinforcement Learning algorithm (Sutton and Barto, 1998) with linear function approximation to train the policies. Figure

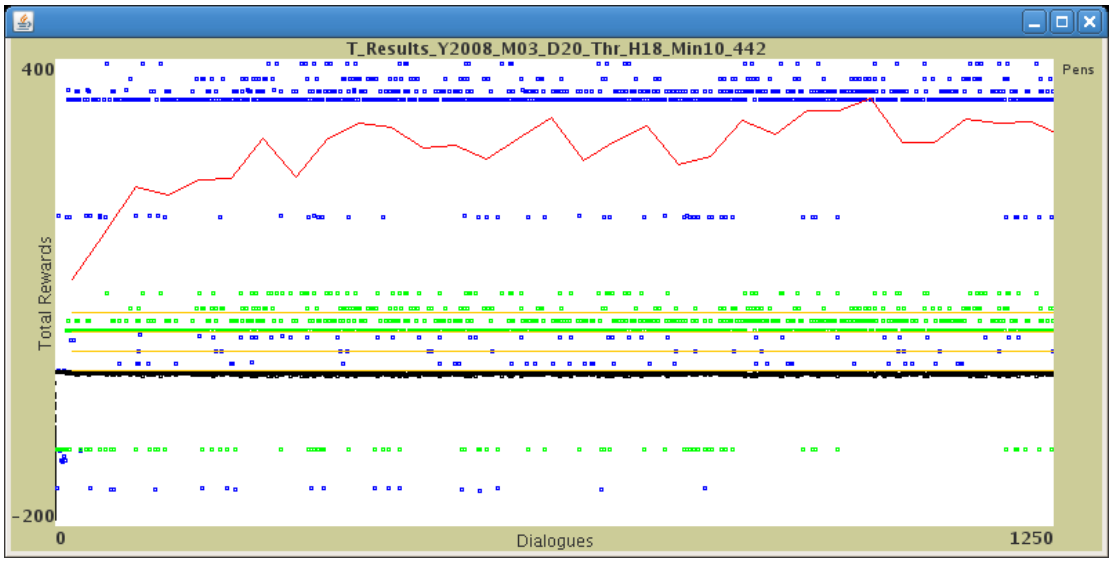


Figure 2: Training the adaptive NLG policy (red line = average dialogue reward)[1]

2 shows learning for the adaptive NLG problem[1].

Here we see that after 1250 training dialogues the system has learned to find a high average reward for the combined NLG and DM problem. At the start of training the system explores bad actions in some states, for example the minimum reward gained in early training is -153, obtained by contrasting more than 7 items (-100) when only 1 slot is filled (-50) after 3 system turns (-3). However, by the end of this training run, the system is able to consistently obtain the best possible rewards given the dialogue situation, for example gaining a top reward of 396 for either Contrast or Cluster of appropriate numbers of items (+300), when all slots are confirmed (+100) in system 4 turns (-4). Where no +300 presentation reward is possible (i.e. $i = 1, 7,$ or 8) the system has learned to Cluster or List (when i=1) the items after filling and confirming all slots. A similar graph can be shown for training the Baseline policy.

4.7 Testing

We trained both policies multiple times until convergence (approx. 10K cycles), selected the best policy in each case, and tested them (with stochastic simulated users) for 550 test dialogues each. Table 1

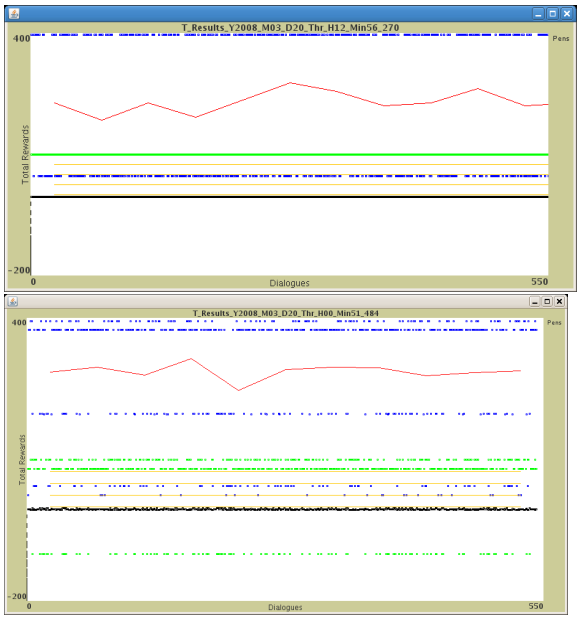


Table 1: Testing the Baseline (top, av. =224.5) and Adaptive (bottom, av. =286.9) NLG policies

[1]In the training/testing graphs red lines show average reward over windows of 50 dialogues, and for each dialogue blue dots show total reward (including NLG reward), black dots show length penalty, and green show completion reward per dialogue.

Policy	Av. Reward	Av. length
Baseline Learned	224.5	4.0
Adaptive NLG Learned	286.9*	4.98

Table 2: Results: learned baseline vs. adaptive NLG policies. ($* = p < 0.001$)

shows the performance of the 2 policies during testing (top= baseline NLG, bottom = adaptive NLG), and the results are presented in table 2.

These results demonstrate a relative increase in reward of 27.8% for the adaptive NLG system. The adaptive NLG system has learned fine-grained local trade-offs for its NLG decisions, which are not available to the baseline system.

So what has been learned? Here is an example dialogue with the adaptive NLG system:
`System`: How can I help you? `(greet)`
`User:` I want a cheap chinese restaurant. (2 slots filled, 2 database items returned)
`System:` Ok. The Golden Wok is cheap and central, and the Noodle bar is cheap but in the south of the city `(Contrast)`

Here we can see that the adaptive NLG policy can decide to present information when it is particularly advantageous, even when the information gathering part of the system is not complete. The Baseline learns a similar policy, but is not sensitive to DB hits when choosing *how* to present the information.

5 Summary and Future Directions

This paper demonstrates a new data-driven method where the NLG components of dialogue systems can be automatically trained and globally optimized before deployment.

We surveyed standard approaches to NLG, and described general advantages offered by statistical planning models together with solution methods such as Reinforcement Learning. We gave a brief description of MDP models. In section 4 we cast a standard NLG problem as an MDP, defining the state space, action set, and reward function. We saw how Reinforcement Learning can be used to solve this NLG problem at the same time as optimizing dialogue management. We then evaluated the adaptive NLG policy versus a learned RL-majority baseline. The results showed a significant relative increase in reward of 27.8% for the adaptive NLG system. When given a reward signal that provides feedback on content structuring choices (List, Contrast, Cluster) the system learns to avoid bad decisions (e.g. listing lots of items, clustering small numbers of items, contrasting too few or too many items) and to choose the best NLG option available depending on the number of database items returned by the system at any time.

This demonstrates that our approach brings a number of theoretical and practical benefits such as fine-grained adaptation, and automatic optimization. Future challenges include modelling the hierarchical structure of NLG problems using additional hierarchical MDPs, and modelling complex effects of NLG choices on dialogue context using larger feature sets.

Many other NLG decisions could be approached in this way. By using MDPs to represent other NLG problems we can move to a situation where determination of the best lexical items and referring expressions to use in a system utterance, as well as the best syntactic structure and intonation pattern, are all determined by learned strategies, developed by reward-driven learning based on real data. Reinforcement Learning could also be applied to decisions of when and how to use anaphora and ellipsis.

Overall, this leads us to propose a new development cycle for NLG, whereby the more adaptive NLG components of new dialogue systems can be automatically trained and optimized before deployment, and can then be allowed to adapt online to user feedback (through continued monitoring of rewards). Moreover, due to the use of state generalization techniques such as function approximation, NLG will even be possible in previously unseen and unplanned-for situations.

An open question for this type of model is how to develop "good" user simulations that are sensitive to system NLG choices (Janarthanam and Lemon, 2008). Another important topic is how the classifier-based learning techniques of "trainable" NLG (Barzilay and Lapata, 2005; Duboue and McKeown, 2003; Stent et al., 2004; Walker et al., 2007) can be integrated with the MDP approach proposed here. Other avenues to explore are how interactive alignment (Garrod and Pickering, 2001) and semantic coordination in dialogue (Larsson, 2007)

can be modelled in this framework.

Acknowledgments

The research leading to these results has received funding from the EPSRC (project no. EP/E019501/1) and from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (CLASSiC project `www.classic-project.org`)

References

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of EMNLP*.

Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952.

Vera Demberg and Johanna D. Moore. 2006. Information presentation in spoken dialogue systems. In *Proceedings of EACL*.

Pablo A. Duboue and Kathleen R. McKeown. 2003. Statistical acquisition of content selection rules for natural language generation. In *Proceedings of EMNLP*.

Simon Garrod and Martin Pickering. 2001. Toward a mechanistic psychology of dialogue: The interactive alignment model. In *Proceedings of BI-dialog*.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *Proceedings of Interspeech/ICSLP*, pages 1065–1068.

Amy Isard, Jon Oberlander, Ion Androutsopoulos, and Colin Matheson. 2003. Speaking the users' languages. *IEEE Intelligent Systems Magazine*, 18(1):40–45.

Srinivasan Janarthanam and Oliver Lemon. 2008. User simulations for online adaptation and knowledge-alignment in Troubleshooting dialogue systems. In *Proceedings of SEMdial*.

Alexander Koller and Matthew Stone. 2007. Sentence generation as planning. In *Proceedings of ACL*.

Staffan Larsson. 2007. Coordinating on ad-hoc semantic systems in dialogue. In *Proceedings of DECALOG*.

Oliver Lemon and Xingkun Liu. 2007. Dialogue policy learning for combinations of noise and user simulation: transfer results. In *SIGdial*.

E. Levin and R. Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies. In *Proceedings of Eurospeech*.

Johanna Moore, Mary Ellen Foster, Oliver Lemon, and Michael White. 2004. Generating tailored, comparative descriptions in spoken dialogue. In *Proc. FLAIRS*.

Alice Oh and Alexander Rudnicky. 2002. Stochastic natural language generation for spoken dialog systems. *Computer, Speech & Language*, 16(3/4):387–407.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. CUP.

Verena Rieser and Oliver Lemon. 2008. Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz data: Bootstrapping and Evaluation. In *Proceedings of ACL*, page (to appear).

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research (JAIR)*.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Association for Computational Linguistics*.

Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. 2003. Microplanning with communicative intentions: the SPUD system. *Computational Intelligence*, 19(4):311–381.

Richard Sutton and Andrew Barto. 1998. *Reinforcement Learning*. MIT Press.

Marilyn A. Walker, Jeanne C. Fromer, and Shrikanth Narayanan. 1998. Learning optimal dialogue strategies: a case study of a spoken dialogue agent for email. In *Proceedings of ACL*.

Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. 2000. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*, 6(3).

M. Walker, O. Rambow, and M. Rogati. 2001. Spot: A trainable sentence planner. In *In Proc. of the NAACL,*.

Marilyn Walker, S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy. 2004. User tailored generation in the match multimodal dialogue system. *Cognitive Science*, 28:811–840.

Marilyn Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.

Steve Young. 2000. Probabilistic methods in spoken dialogue systems. *Philosophical Transactions of the Royal Society (Series A)*, 358(1769):1389–1402.

Taking Fingerprints of Speech-and-Gesture Ensembles

Approaching Empirical Evidence of Intrapersonal Alignment in Multimodal Communication

Andy Lücking CRC 673, B1
Alexander Mehler CRC 673, A3, X1
Peter Menke* CRC 673, X1
Bielefeld University, CRC 673 "Alignment in Communication"
{Andy.Luecking,Alexander.Mehler,Peter.Menke}@uni-bielefeld.de

Abstract

Co-occurring speech and gestures of natural language dialogues composes into meaning units, that is, they jointly describe discourse referents. We start from the idea that interlocutors tend to re-use this cross-modal information units if the discourse referent is referred to again: co-occurring speech and gesture are assumed to "align into" *bimodal ensembles* (BMEs). We further hypothesize that due to principles of dialogical economy interlocutors will exploit the impact of a BME's gesture to shorten its linguistic part of that BME. If this hypothesis is right, we expect that the words in multimodal communication exhibit a different frequency distribution from words in written texts, whose frequency distribution is known to obey Zipf's law. This hypothesis is tested for 24 direction-giving dialogues using two different frequency fits, rank frequency distribution and complementary cumulative distribution. According to the first fit, the hypothesis can be confirmed, according to the second one, it has to be rejected. In addition, we also propose a way to measure the strength of cross-modal informational association.

1 Introduction and Reasoning

This article presents some ideas about how to combine text-technological tools and linguistic research in the study of multi-modal dialogue, that is dialogue comprising speech and gesture. The term 'gesture' refers to gesticulations according to Kendon's continuum (Kendon, 1988), that is, 'gesture' is understood as a spontaneous co-verbal hand and arm movement which is linguistically significant and contributes to the narrative. McNeill (1992), alluding to a Peircean trichotomy, distinguishes different types of gesture, namely *deictic* gestures, *iconic* gestures, and *beats*. Beats are rhythmic stresses, deictic gestures are pointings. According to Peirce, icons are representations ("signs"), "whose relation to their objects is a mere community in some quality" (Peirce, 1867). That is, icons signify due to a certain resemblance between signifier and signified.

However, 'icon' is an "umbrella term" (cf. (Eco, 1976)) that covers a variety of different signifying methods. (Müller, 1998), drawing on the work of (Wundt, 1911), sets up a more fine-grained classification of gestures according to the distinction of four *modes of representation* on the ground of what the hands *do: Agieren* (Acting), *Modellieren* (Modelling), *Zeichnen* (Drawing), and *Repräsentieren* (Representing).

Ancient rhetoric already emphasizes the rhetoric connection between speech and gesture (Quintilian, 1st century; Maier-Eichhorn, 1989). In modern times, most notable Kendon (1980) claims that verbal utterances together with simultaneous accompanying deictic and iconic gestures coheres into single meaning units. However, it is not yet clear how the mechanism that binds together the two communication channels should be modelled – is it functional application (Rieser, 2004), rhetorical relation

*Authors' names are given in alphabetical order.

(Lücking et al., 2006; Lascarides and Stone, 2006) or something else? For the time being the pair of gesture and affiliated speech should be construed as an informational wholeness tied together by some kind of synchronicity principle (Jung, 1971). Take for instance an example from the study described in Section 2, where a subject is talking about one of two churches on a square which are, amongst others, distinguished by the type of their roofs:

(1) rechts die hat so'n [Giebel]
the one to the right it has such a [gable]
∧-shaped gesture synchronous to bracketed speech

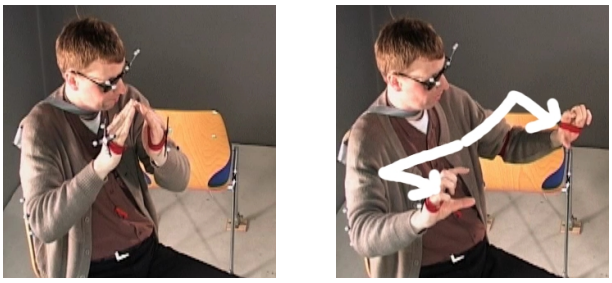
The gesture from (1), which is displayed as Figure 1(a), is a *Posturing* gesture according to the modes of representation scheme introduced below. We assume that for the period of the dialogue the gesture gets associated with its accompanying speech,[1] or, as we will call it hereafter: The bracketed portion of the linguistic utterance together with the accompanying speech constitutes a *bimodal ensemble* (BME).

The linguistic part of a BME may comprise more than single words, as is illustrated in (2), where the subject talks about a chapel that is located within the "punch" of a surrounding "⊔"-shaped hedge, as indicated by a *Shaping* gesture (see Figure 1(b)).

(2) die hat ['ne grüne Hecke drumherum]
it has [a green hedgerow around it]
⊔-shaped gesture synchronous to bracketed speech

There is some discussion about the informational relation between speech and gesture: Is it redundancy or complementarity? (Cassell and Prevost, 1996; Bergmann and Kopp, 2006) We will, however, bypass this issue since our concern is purely quantitative: The linguistic part of BMEs is the input for the frequency distribution analysis given in Section 3.

On a more abstract level, a BME is an assemblage comprising a set of parts of speech (classes of words) and a representation technique (class of gesture, e.g. *Shaping*). BMEs conceived this way enter into the determination of the Hartley information (Klir and Folger, 1988) (see Section 3).



(a) Gable (b) Hedgerow

Figure 1: Two sample gestures.

The fusion of speech and gesture into a BME in dialogue is a precondition to the investigation pursued in this article. We investigate a **hypothesis concerning bimodal ensembles**: The use of gesture facilitates a merely partial recurrence or a paraphrase of the linguistic material of a BME. Since there is not yet data directed to and annotated for speech-and-gesture coupling over the time-course or a dialogue, we approach this issue by means of an indirect measuring. Think, for example, of an ensemble $e = (xy, g)$ manifested by some linguistic material xy in conjunction with a gesture g. The interlocutors of D may manifest e later on by the parts x, y of xy (or maybe even by some unit z which is sense-related to x, y or xy). The reason is that the simultaneously produced gesture g allows for correctly disambiguating the shortened or otherwise modified linguistic manifestation of the ensemble e.[2] For an illustration take the sample utterance (2). The BME (*'ne grüne Hecke drumherum*, ⊔) might get shortened to:

(3) die hat ['ne Hecke]
it has [a hedgerow]
⊔-shaped gesture

To give an example for sense related substitution: The word *Giebel/gable* from (1) might be replaced by the hyponym *Pediment*:

(4) die hat so'n [Pediment]
it has such a [gable]

[1] Most presumably, the association is established by some grounding mechanism (see for instance (Clark and Schaefer, 1989)), but we will not pursue this issue further here.

[2] It may also be the case that interlocutors reduce motor effort and produce simplified gestures. But this is a different story.

$\wedge$-*shaped gesture*

The described mechanisms leave an option to express the same concept in dialogical communication. Thus, any frequent usage of this method of reducing communication effort has an impact on the frequency distribution of lexical units within D: the same concept denoted by e is alternatively manifested by xy, x, y, z, z', z'', ... (Remember that z, z', z'' are sense related to x, y, or xy.) As this method of lexical choice is out of reach in written communication we expect an impact of using gestures on the frequency distributions of lexical units in dialogues.

Note that this argumentation presupposes that there is a usage-chain in dialogue D from the BME e to its shortening later on in D.

Note further that we do not expect this effect on the level of highly frequent words which, as expected, consist of function words and therefore rarely count as linguistic manifestations of bimodal ensembles.

The next section gives a brief overview of the study that underlies the data our investigation is based on. It also introduces the gestural representation techniques that enter into the determination of the Hartley information. The measuring procedures and its results are given in Section 3.

2 Experimental Study

Iconic gesturing is inherently spatial (Alibali, 2005). A kind of setting that has proved to elicit spatial discourse is the description of routes (Denis, 1997). Accordingly, the empirical data of our research consists of direction giving dialogues. The dialogues are about city tours one of the interlocutors has made in a town presented in a Virtual Reality environment (Kopp et al., 2008). Thus, our empirical study comprises two phases: At first, a participant undertakes a "bus ride" in a virtual town, see Figure 2(a) for an illustration. The sight-seeing tour passes five objects of interest, namely an abstract sculpture, a city hall, a church square with two churches, a chapel and a fountain. Subsequently, the first participant, called Router (R), has to explain to a second participant who does not know the virtual town which route he has driven and what landmarks he has seen. In order to elicit an elaborate spatial discourse the second participant, Follower (F), was made to believe that he will have to find the route through the virtual town and to identify all landmarks. Splitting up the virtual sight-seeing tour in a route and a landmark part, different types of spatial communication will come up, namely giving directions and describing shapes. Both are good candidates for iconic depiction.

In view of the frequency analysis to come, the employment of a virtual stimulus is a precondition for the inter-participant comparability of linguistic and non-verbal data, since it assures that all participants talk about the same thing.

2.1 Annotation

Annotation layers divide naturally into two different partitions, the one relating to speech the other relating to gestures. Speech transcription has been made using Praat[3] and has been done orthographically, i.e., on the level of words. Part of speech information is added automatically by means of POS-tagging (Gleim et al., 2007).

For gesture annotation, we delimit the gesture's semantic phase known as *stroke* (McNeill, 1992). Each stroke has been assigned a mode of representation. We have extended and terminologically modified Müller's set of representation modes in order to adjust it to the specific needs of route descriptions. Gesture has been annotated using the multimedia annotation software Elan[4]. The representation techniques we recognize are itemized and briefly commented upon in the following list.

Shaping The hands are sliding on the surface of a virtual object in gesture space, a shape emerges.

Sizing A configuration of hands or fingers that indicate a certain distance or size is called Sizing.

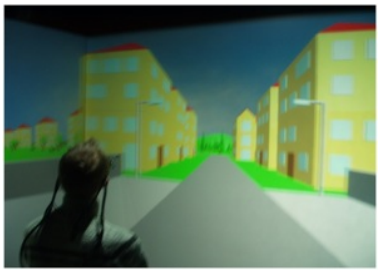
Posturing The hand (or both hands) represent an object involved in the described situation.

Drawing A single finger or the hand is used as a drawing tool to sketch an outline in the gesture space.

Pantomime The usage of an object or an action is displayed by imitation. Note that Pantomime,

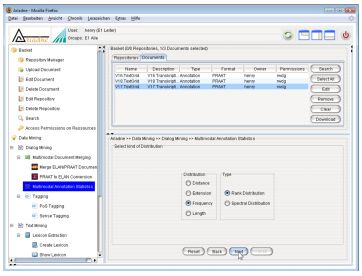
[3] www.fon.hum.uva.nl/praat

[4] www.lat-mpi.eu/tools/elan



(a) Virtual bus ride

(b) Route description



(c) Ariadne system

Figure 2: Virtual environment stimulus and subsequent dialogue: the Data are managed in the Ariadne system.

in contrast to the other gesture practices, makes the gesturer himself a part of the depiction, not just his hands or arms.

Indexing A deictic gesture that singles out a point in the gesture space which thereby gets "semantically loaded", e.g., becomes a proxy for an object of the narrative.

Grasping If the hand touches or holds an object, but does not shape its body, then a Grasping-gesture is performed.

Counting If the fingers are used to enumerate things. Gestural counting can be seen as an iconic representation of a tally sheet.

Hedging Sometimes a wiggling or shrugging movement is used in order to depict uncertainty. We call this metaphoric gesture method 'Hedging'.

In sum, there are 25 direction-giving dyads with a total of 4961 gestures and 39.435 words.

Our multimodal dialogue data are stored, retrieved, transformed, and statistically explored within the Ariadne system (Gleim et al., 2007) which is used as an *Alignment Corpus Management System* (ACMS) – see the screen shot displayed as Figure 2(c).

2.2 Reliability

Since the classification of gestures in terms of representation modes is interpretive data, it is questionable whether it is reproducible (Krippendorff, 1980). Our evaluation of gesture classification data follows the discussion in (Stegmann and Lücking, 2005). A sample of gestures large enough to test for the reasonable agreement level of 70% with an α-error of 0.05 and a β-error of 0.85 (set in the run-up to the reliability study) has been classified by three expert annotators. The resulting first-order agreement coefficient AC_1 (Gwet, 2001) is 0.784. It's confidence interval is $(0.758, 0.81)$, so that the probability for agreement on gestures' representation modes – given that the agreement is not due to chance – is significantly greater than 75%.

3 Measuring Procedure and Results

Our starting point of indirectly measuring an impact of gestures on the choice of lexical units is Zipf's law (Zipf, 1972) which we denote as follows (Adamic, 2000):

$$n \sim r^{-\gamma} \tag{1}$$

n is the frequency of the rth most frequent word in the given text (or dialogue) for which Model (1) is fitted. Roughly, Zipf and related studies show that $\gamma \sim 1$ for written texts (Rapoport, 1982; Tuldava, 1998). Taking this as a reference value we expect – according to our hypothesis – a lower value of γ in the case of dialogical communication, that is, a flatter straight line which results from a log-log plot of the Rank Frequency Model (1). Note that $-\gamma$ is the slope of that line. Look, for example, at Figure 3(a), where we have fitted the power law $Cx^{-\gamma}$ to the *Rank Frequency Distribution* (RFD) of lexical units used by some interlocutor in a dialogue from our corpus. That is, the first rank is the one of the

most frequent word, the second the one of the second most frequent word, and so on till we finally reach the ranks of *hapax legomena*. Fitting this empirical curve and plotting the result in a log-log plot we see that $\gamma = .678$ while the adjusted coefficient of determination $\bar{R}^2$ equals .9674. This indicates a good fit.[5] This result is in support of our hypothesis of a gesture-based impact on lexical choices – it does not falsify the hypothesis about the existence of this impact: as the exponent is smaller than one, the curve is flatter than suggested by the results derived from written texts.

However, according to (Newman, 2005) fittings change for the better by operating on the *Complementary Cumulative Distribution* (CCD), that is, on the probability function $P(X \geq x)$ of words which occur at least x times. In the case of our example the results of fitting to the CCD derived from the corresponding RFD are shown in Figure 3(b): Now, $\gamma = 1.145$ and $\bar{R}^2 = .9937$ what indicates a slightly better fit. *Can these two measurements be compared?* According to (Adamic, 2000) the exponent γ of a Zipfian RFD corresponding to a given CCD with exponent β is computed by $\gamma = 1/\beta$ – in the present case we achieve $\gamma \sim .873$.

That is, relying on the CCD which gives a better fit than the previously observed RFD and deriving the exponent of a RFD – which corresponds to the latter CCD on the same level of goodness of fitting – the absolute value of the exponent is raised ($.873 > .678$). This is what we actually observe in nearly all cases of our corpus of 24 dialogues.[6] The corresponding box plots of the 24 exponents γ and the corresponding determination coefficients $\bar{R}^2$ are shown in Figure 4: Not only are the absolute values of the exponents of the power laws fitted to the corresponding CCDs higher than the one of the primarily observed RFDs. More important is the observation of remarkably higher values of $\bar{R}^2$ – that is, as indicated by (Newman, 2005), CCDs are more reliable reference points of power law fitting. Thus, we additionally derive – according to the approach of (Adamic, 2000) – the exponents of those rank frequency distributions which correspond to the latter CCDs on the same level of goodness of fitting. As a result we see that we get on average higher values than in the case of the primarily observed RFDs (cf. Figure 4(c)). Moreover, the newly derived values disperse around 1 and are, therefore, in a good neighborhood of those values which were observed by Zipf. In this sense, our results do not indicate a difference between written and dialogical communication – at least under the regime of our experimental setting. Following this line of argumentation, there is *no* effect on the frequency distribution of lexical units. This hypothesis is only upheld by referring to the fittings of the left part of Figure 4 – however at the price of relying on worse fittings.

As our distribution analysis does not shed much light on the existence of bimodal ensembles we now compute a measure of interactivity between selections on the lexical and gestural layer. Such cross-modal selections are called interactive if, for example, the selection of lexical units constrains the selection of gestural units, that is, if there is a tendency of co-occurrence among lexical and gestural units. If we could measure such a tendency, this could be interpreted as a support of our hypothesis about the existence of bimodal ensembles. As we will see, this is not achieved.

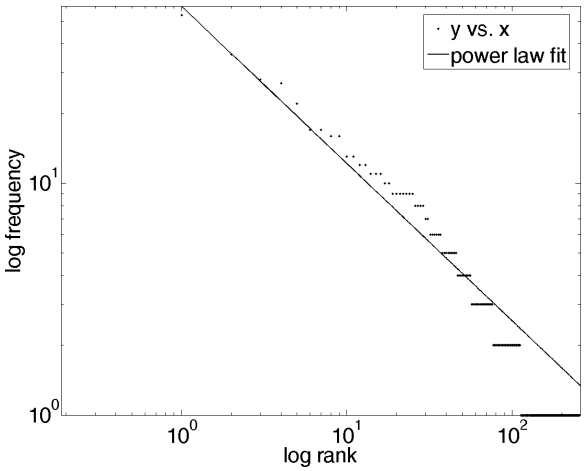
In order to get a first measure of the interaction of cross-modal selections we compute the information transmission between selecting from the set of parts of speech X and the set of representation techniques Y. Generally speaking, the information transmission between n sets $X_1, \ldots, X_n$ is defined as follows (Klir and Folger, 1988):

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{n} I(X_i) - I(X_1, \ldots, X_n) \qquad (2)$$

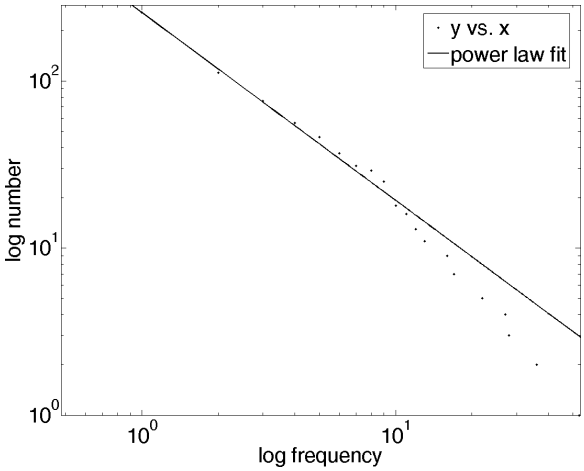
where $I(X) = \log_2 |X|$ is the simple Hartley information of X (cf. (Klir and Folger, 1988) for the details of this and related definitions), $I(X,Y) = \log_2 |R|, R \subseteq X \times Y$, is the joint (Hartley) information. In our case R is the set of all co-articulated parts of speech and representation techniques: Generally speaking, the sets $X_1, \ldots, X_n$ are called *non-interactive* if $T(X_1, \ldots, X_n) = 0$, otherwise we observe that $T(X_1, \ldots, X_n) > 0$. Note that $T(X_1, \ldots, X_n) = 0$ if and only if $R = X_1 \times \ldots \times X_n$. In this case, any selection from set $X_i, i \in \{1, \ldots, n\}$,

[5] The adjusted coefficient of determination is a measure of goodness of fitting: the nearer its value to 1, the better the fit.

[6] Note that we deleted one dialogue from the corpus because of too many uncertain annotations.



(a) Results of fitting to the *Rank Frequency Distribution* (RFD – Zipfian scenario).



(b) Results of fitting to the *Complementary Cumulative Distribution* (CCD) derived from the latter RFD.

Figure 3: Two sample power law fittings of the frequency distribution of lexical units of a single interlocutor (in the role of the *router*). In both cases, the model $y = Cx^{-\gamma}$ is used.

may be combined with any selection from any other set $X_j, j \in \{1,\ldots,n\} \setminus \{i\}$. As the range of values of T is not limited, we standardize it as follows:

$$\hat{T}(X_1,\ldots,X_n) = \frac{T(X_1,\ldots,X_n)}{\sum_{i=1}^{n} I(X_i)} \in [0,1] \qquad (3)$$

Now, we see that for $\hat{T}(X_1,\ldots,X_n) \ll 1$ the sets $X_1,\ldots,X_n$ tend to be non-interactive, while they tend to be interactive if in contrast to this $\hat{T}(X_1,\ldots,X_n) \gg 0$. In other words: 0 indicates minimal and 1 maximal interactivity. In Figure 5 we report the results of measuring the interaction between the selection of parts of speech and of gestural practices by 24 interlocutors in 24 dialogues. Obviously, the sets are far from being interactive according to this measure of interactivity (which measures on an ordinal scale). However, as we do not yet know anything about expected values of such an interaction among elements of different modes in multimodal communication, we hesitate to value this as a falsification of our starting hypothesis. Anyhow, this hypothesis is not supported by both of our measurements, neither on the level of lexical distributions nor on the level of interactions of cross-modal choices. If we rely on the classical operation of rank frequency distribution, our hypothesis is not falsified. However, if we use the CCD we get a hint that there is no distributional difference.

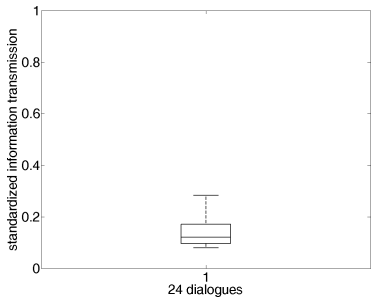
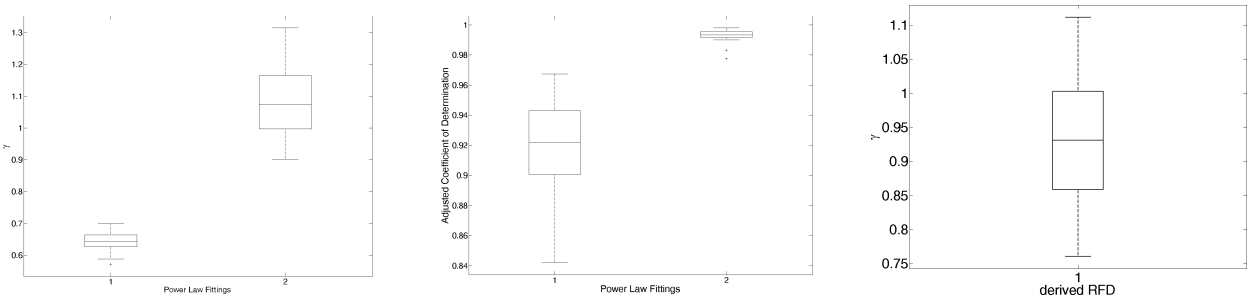


Figure 5: The distribution of information transmission between the selection of parts of speech and gestural practices by interlocutors in 24 dialogues.

4 Conclusion

One reason for the rather ambivalent result might be that its underlying presupposition does not hold. Ambivalent means that the rejection of the hypothesis depends on whether the fit is based on chosing the rank frequency distribution or the complementary cumulative distribution.

Recall from Section 1 that a BME e leaves a frequency distributional fingerprint only if there is a usage-chain connecting first occurrences of a fully specified e to subsequent shortened manifestations. The progressive rhematic structure of the direction-



(a) Exponent values

(b) Adjusted coefficients of determination

(c) Range of derived exponents of rank frequency distributions

Figure 4: Box plots of the exponent values (a), the corresponding adjusted coefficients of determination (b), and of the range of derived exponents of rank frequency distributions which correspond to the primarily observed complementary cumulative distributions (c) of all 24 dialogues of our corpus. The first column of both the (a) and the (b) sub-figures denotes the rank frequency model while the second column denotes the complementary cumulative model.

giving dialogues might block the establishment of a usage-chain for a certain BME *e*, leaving *e* an merely ephemeral phenomenon.[7]

As exposed in the preceding section, the rejection or affirmation of the hypothesis investigated in our analysis partly depends on "baseline values" for the different measuring procedures. Even if we cannot maintain our working hypothesis – and we have been very careful not to overstate our results, cf. Section 3 – analyses like the one carried out make up the pieces of the puzzle needed in order for a more comprehensive exploration of multimodal data. If BMEs indeed leave fingerprints that are measurable in the way explored in this article, this result clearly has an impact on cognitive theories, for instance theories of speech-and-gesture production. If there is an intra-personal alignment of words and gesture during a dialogue, the production of units on the respective modalities interacts. That is, empirical, quantitative research like the one presented here might help to collect evidence for or against different views of production processes as developed by (McNeill and Duncan, 2000; Kita and Özyürek, 2003; de Ruiter, 2000; Krauss et al., 2000).

Acknowledgements. Thanks to Hannes Rieser, Stefan Kopp, Kirsten Bergmann, and Florian Hahn for discussions and/or data annotation. We also want to thank the anonymous reviewers whose critique helped to clarify the ideas presented here. This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication".

[7]Note that the focus of our line of reasoning is a frequency *distribution*, *not* the *frequency* of (gesture accompanying) words. Thus our result does not contradict studies which prove the latter.

References

Lada A. Adamic. 2000. Zipf, power-law, Pareto – a ranking tutorial.

Martha W. Alibali. 2005. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, 5:307–331.

Kirsten Bergmann and Stefan Kopp. 2006. Verbal or visual? how information is distributed across speech and gesture in spatial dialog. In David Schlangen and Raquel Fernández, editors, *brandial'06 – Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 90–97, Potsdam.

Justin Cassell and Scott Prevost. 1996. Distribution of semantic features across speech and gesture by humans and computers. In *Proceedings of the Workshop on Integration of Gesture in Language and Speech*.

Herbert H Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Jan Peter de Ruiter. 2000. The production of gesture and speech. In David McNeill, editor, *Language and gesture*. Cambridge University Press.

Michel Denis. 1997. The description of routes: A cognitive appraoch to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458.

Umberto Eco. 1976. *A Theory of Semiotics*. Indiana University Press, Bloomington.

Rüdiger Gleim, Alexander Mehler, and Hans-Jürgen Eikmeyer. 2007. Representing and maintaining large corpora. In *Proceedings of the Corpus Linguistics 2007 Conference, Birmingham (UK)*.

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, Gaithersburg, MD.

Carl Gustav Jung. 1971. Synchronizität als ein Prinzip akausaler Zusammenhänge. In *Die Dynamik des Unbewussten*, volume 8 of *Gesammelte Werke*, chapter XVIII, pages 475–577. Walter-Verlag, Olten and Freiburg im Breisgau. First published in *Naturerklärung und Psyche* (1952).

Adam Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Mary Ritchie Key, editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. Mouton Publishers, The Hague.

Adam Kendon. 1988. How gestures can become like words. In F. Poyatos, editor, *Cross-cultural Perspectives in Non-Verbal Communication*, pages 131–141. Hogrefe, Toronto.

Sotaro Kita and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32.

George J. Klir and Tina A. Folger. 1988. *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall, Englewood.

Stefan Kopp, Kirsten Bergmann, and Ipke Wachsmuth. 2008. Multimodal communication from multimodal thinking—towards an integrated model of speech and gesture production. *International Journal of Semantic Computing*. To be published.

Robert M. Krauss, Yihsiu Chen, and Rebecca F. Gottesman. 2000. Lexical gestures and lexical access: A process model. In David McNeill, editor, *Language and gesture*. Cambridge University Press.

Klaus Krippendorff. 1980. *Content analysis*. SAGE Publications, Beverly Hills.

Alex Lascarides and Matthew Stone. 2006. Formal semantics of iconic gesture. In David Schlangen and Raquel Fernández, editors, *brandial'06 Proceedings*, pages 64–71, Potsdam.

Andy Lücking, Hannes Rieser, and Marc Staudacher. 2006. Multi-modal integration for gesture and speech. In David Schlangen and Raquel Fernández, editors, *brandial'06 Proceedings*, pages 106–113, Potsdam.

Ursula Maier-Eichhorn. 1989. *Die Gestikulation in Quintilians Rhetorik*. Peter Lang, Frankfurt am Main.

David McNeill and Sunsan Duncan. 2000. Growth points in thinking-for-speaking. In David McNeill, editor, *Language and gesture*. Cambridge University Press.

David McNeill. 1992. *Hand and Mind—What Gestures Reveal about Thought*. Chicago University Press, Chicago.

Cornelia Müller. 1998. *Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich*, volume 1 of *Körper – Kultur – Kommunikation*. Berlin Verlag, Berlin.

M. E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351.

Charles Sanders Peirce. 1867. On a new list of categories. In *Proceedings of the American Academy of Arts and Sciences Series*, volume 7, pages 287–298.

Marcus Fabius Quintilian. 1st century. *Institutio Oratoria*.

Anatol Rapoport. 1982. Zipf's law re-visited. In H. Guiter and M. V. Arapov, editors, *Studies on Zipf's Law*, pages 1–28. Brockmeyer, Bochum.

Hannes Rieser. 2004. Pointing in dialogue. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04 Proceedings*, pages 93–100, Barcelona.

Jens Stegmann and Andy Lücking. 2005. Assessing reliability on annotations (1): Theoretical considerations. Technical Report 2, CRC 360, Universität Bielefeld.

Juhan Tuldava. 1998. *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Wissenschaftlicher Verlag, Trier.

Wilhelm Wundt. 1911. *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus und Sitte*, volume I: Die Sprache. Erster Teil. Wilhelm Engelmann, Leipzig.

George K. Zipf. 1972. *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Hafner Publishing Company, New York.

Multimodal Reference in Dialogue: Towards a Balanced Corpus

Paul Piwek
Centre for Research in Computing
The Open University, UK
p.piwek@open.ac.uk

Ielka van der Sluis
Computer Science
Trinity College Dublin, Ireland
ielka.vandersluis@cs.tcd.ie

Albert Gatt
Computing Science
University of Aberdeen, UK
a.gatt@abdn.ac.uk

Adrian Bangerter
Institut de Psychologie du Travail et des Organisations
University of Neuchâtel, Switzerland
adrian.bangerter@unine.ch

Introduction

Generation of Referring Expressions (GRE), e.g., Dale and Reiter (1995), is one of the core tasks of Natural Language Generation (NLG) systems. Usually it is formulated as an identification problem: given a domain representing entities and their properties, construct a referring expression for a target referent or set of target referents which singles it out from its distractors. Recently, researchers in this area have turned their attention to *multimodal referring acts*, in particular, the interaction between the two modalities of *pointing* and *describing* – e.g., Kranstedt et al. (2006), Piwek (2007), and Van der Sluis and Krahmer (2007). Additionally, psycholinguistic work is increasingly investigating the conditions governing the use of pointing gestures as part of referring acts in *dialogue*, opposed to *monologue*. Here, we present the design of an experiment on multimodal reference in two-party dialogue. The purpose of the experiment is to create a corpus that can inform the development of multimodal GRE algorithms.

Collecting a Balanced Corpus

We have paid specific attention to balancing the corpus: the conditions under which references were elicited correspond to experimental variables that are counter-balanced. The use of a dialogue setting will allow us to investigate both the speaker/generator's and hearer/reader's point of view, with potentially useful data on such factors as alignment and entrainment, and the nature of collaboration or negotiation, topics of much debate in the psycholinguistic literature (Pickering and Garrod, 2004).

In our setup for collecting dialogues, a director and a follower are talking about a map that is situated on the wall in front of them, henceforth the *shared map*. Both can interact freely using speech and gesture, without touching the shared map or standing up. Each also has a private copy of the map; the director's copy has an itinerary on it, and her task is to communicate the itinerary to the follower. The follower needs to reproduce the itinerary on his private copy. The rules of for the interaction were as follows:

- Since this is a conversation, the follower is free to interrupt the director and ask for any clarification s/he thinks is necessary.
- Both participants are free to indicate landmarks or parts of the shared map to their partner in any way they like.
- Both participants are not permitted to show their partner their private map at any point. They can only discuss the shared map.
- Both participants must remain seated throughout the experiment.

While this task resembles the MapTask experiments (Anderson et al., 1991), the latter manipulated mismatches between features on the director and follower map, phonological properties of feature labels on maps, familiarity of participants with each other, and eye contact between participants. The current experiment systematically manipulates target size, colour, cardinality, prior reference and domain focus, in a balanced design. Though this arguably leads to a certain degree of artificiality in the conversational setting, the balance would not be easy to obtain in an uncontrolled setting or with off-the-shelf materials like real maps. Further properties of

our experiment that distinguish it from the MapTask are: (1) objects in the visual domains are not named, so that participants need to produce their own referring expressions, (2) the participants are always able to see each other; (3) the participants are allowed to include pointing gestures in their referring expressions.

Four maps were constructed, consisting of simple geometrical landmarks (ovals or squares). Two of the maps (one each for ovals and squares) have *group* landmarks, whereas the other two have singletons. Objects differ in their size (large, medium, small) and colour (red, blue, green). Each dyad in the experiment discusses all four maps. Per dyad, the participants switch director/follower roles after each map. The order in which dyads discuss maps is counter balanced across dyads. There are four independent variables in this experiment:

- **Cardinality** The target destinations in the itineraries are either singleton sets or sets of 5 objects that have the same attributes (e.g., all green squares)
- **Visual Attributes:** Targets on the itinerary differ from their distractors – the objects in their immediate vicinity (the 'focus area') – in colour, or in size, or in both colour and size. The focus area is defined as the set of objects immediately surrounding a target.
- **Prior reference:** Some of the targets are visited twice in the itinerary.
- **Shift of domain focus:** Targets are located near to or far away from the previous target. If two targets t_1 and t_2 are in the *near* condition, then t_1 is one of the distractors of t_2 and vice versa.

Current Status and Further Work

After a pilot of the experiment, data was collected from 22 dyads with the validated setup. Currently, the data is being transcribed, see Figure 1 for an example. Our next task is to annotate the data, focussing on identification of multimodal referring expressions, linking of referring expressions with domain objects (i.e., intended referents) and segmentation of dialogue into episodes spanning the point in time from initiation to successful completion of a target identification. Elsewhere (van der Sluis et al., 2008), we provide information on the hypotheses that we intend to test on the annotated corpus.

128	D	Uh and if you *go straight up* from that you've got five blue ones	D points at the map and moves his finger upwards
129	F	Yeah [*there?*]	D is still pointing F points
130	D	[There] yeah	D is still pointing F is still pointing
131	F	one two three four five	D is still pointing F is still pointing
132	D	Yeah. They're all number three	D is still pointing
133	F	Right. Right.	
134	D	And the five reds just *to the right over*	D points and moves his finger to the right
135	F	And like a kind of *downwards* arrow	D is still pointing F moves his hand upwards
136	D	Arrow yeah they're all number four. Number five. Uh and five is paired with one *with these ones*.	D stops pointing D points
137	F	All right.	

Figure 1: Excerpt from dialogue O17-S33-S34, where *D* = director, *F* = follower and where the brackets indicate overlapping speech and the text in italics indicates approximately the co-duration of gesture and speech

Acknowledgements This work was partly funded by the EPSRC platform grant 'Affecting people with natural language' (EP/E011764/1).

References

A. Anderson, M. Bader, E. Bard, E. Boyle, G.M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson, and R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366.

R. Dale and E. Reiter. 1995. Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8):233–263.

A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. 2006. Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth, editors, *Situated Communication*, pages 155–208. Mouton de Gruiter.

M. Pickering and S. Garrod. 2004. Toward a Mechanistic Psychology of Dialogue. *Behavioural and Brain Sciences*, 27(2):169–226.

P. Piwek. 2007. Modality choice for generation of referring acts: Pointing versus describing. In *Procs of Workshop on Multimodal Output Generation (MOG 2007)*, Aberdeen, January.

I. van der Sluis and E. Krahmer. 2007. Generating multimodal referring expressions. *Discourse Processes*, 44(3):145–174.

I. van der Sluis, P. Piwek, A. Gatt, and A. Bangerter. 2008. Towards a balanced corpus of multimodal referring expressions in dialogue. In *Procs of Symposium on Multimodal Output Generation (MOG 2008)*, Aberdeen, Scotland, April.

Aligned Iconic Gesture in Different Strata of MM* Route-description Dialogue

Hannes Rieser
Bielefeld University
hannes.rieser@uni-bielefeld.de

Abstract

This paper deals mainly with iconic gesture in two-agent route description dialogue and focuses largely on the interface of word semantics and gesture. The modelling tools used come from formal semantics and pragmatics. The empirical background of the study is a partly annotated corpus of ca 5.000 gestures collected in the Bielefeld Speech-and-Gesture-Alignment Corpus (SAGA). The approach taken is entirely new: an interface comprising word meaning and gesture meaning is constructed, the point of contact being the temporal overlap between gesture and speech in the annotated data. Gesture meaning is computed *via* a mapping *rep* from the set of annotation predicates onto a meaning representation. There is a discussion concerning the trade-off between context-free *vs.* context-dependent word meaning and gesture meaning. The interfaced speech-gesture meaning is represented in a dynamic semantics format easily grafted on a formal syntax fragment.

*MM stands for *multi-modal*.

1 Introduction[1]

It is well known that gestures of agents are ubiquitous in dialogue (cf. McNeill (ed. (2000)), Kita (ed. (2003)) but not where it can be placed in dialogue and what then will be its function there. Judged by experience with corpus data and the gesture folklore there is little doubt that there is pointing to objects in context (cf. Rieser (2008)) and that properties such as rectangularity can in a way be indicated by gesture. However, is there something more definite that can be said? As far as we know there has been no work on MM dialogue so far investigating these matters on a more principled basis. Below it will be shown that gestures can go into different structural positions in dialogue, exhibiting different meanings and functions. Even if we rely on a fairly large corpus of multi-modal dialogue, the (Bielefeld University) SAGA corpus elicited in a strictly controlled VR experiment, comprising roughly 5.000 gestures, the evidence presented here cannot be conclusive. There might still be other functions and most plausibly, there are. Nevertheless, we claim that the findings we show and explain are prototypical for natural MM dialogue. So, in section 1 we will provide an overview on structural positions observed for gestures in MM dialogue. Ch. 2 will deal with a binding problem of some sort, namely, how gesture

[1] In this paper only literature is quoted which has been evaluated as relevant for its methodological concerns, which is largely formal theory building. So, some readers might miss their favourite papers. Thanks go to three anonymous reviewers who raised a lot of interesting issues. Some of their arguments are taken up below, space permitting. Sometimes I will refer to a reviewer's (abbr. as rev. *n*'s) remark.

information can be 'bound'[2] to speech information. Ch. 3 will deal with the interface of gesture meaning and verbal meaning, restricted to word meaning, and there will be a brief discussion of the methodological problems with this approach in ch. 4.

2 Overview on Structural Positions Observed for Gesture in Dialogue

As an introduction to the function of gesture in dialogue, we set out with a naïve methodology and provide prototypical speech-gesture occurrences. We might view these as instances of ratings leading to systematic annotation, i.e. we first do speech-gesture pairings in a naïve way; as a consequence, the total meaning of speech plus gesture is given in the short descriptions. Of course, the coordination of speech and gesture information is a major *explicandum* of this paper, so this introductory perspective will be given up in sections 2-4, where the ontological status of speech meaning and gesture meaning is discussed and the speech-gesture interface is the central issue. The stills in fig. 1 below show the stroke positions of iconic gestures; it should be kept in mind that gestures are incomplete and even non-standard in various ways and provide partial information at best. So, in interpreting gestures we have to assume top-down Gestaltist processes at work. (a) is an oval gesture accompanying the description of a sculpture indicating part of the concrete basis for the sculpture, (b) presents a gesture indicating the two towers of a church, in (c) the route follower imitates the router's gesture indicating the U-shape of the town hall, (d) has an other-correction carried out by a router's gesture, (e) has a two-handed gesture which depicts a situation containing a chapel and a tree. (a) and (b) are routers gestures, (d) has interaction resting on gestures functioning like turns. Fig. 2 gives a summary of these findings, indicating the various functions of gestures.The data in Fig. 1 are related to Fig. 2 as follows: The gesture in still (a) is related to word semantics, the one in (b) to the semantics of an NP-constituent, in (c) a gesture goes proxy for a propositional content which gets acknowledged, (d) shows that a gesture is used in a next turn repair, in (e), finally, the right hand models a tree while the left hand indicates the location of the tree beside a chapel.

The example 1 discussed below (cf. fig. 4) will deal in some detail with the extension of word meaning by gestural meaning.

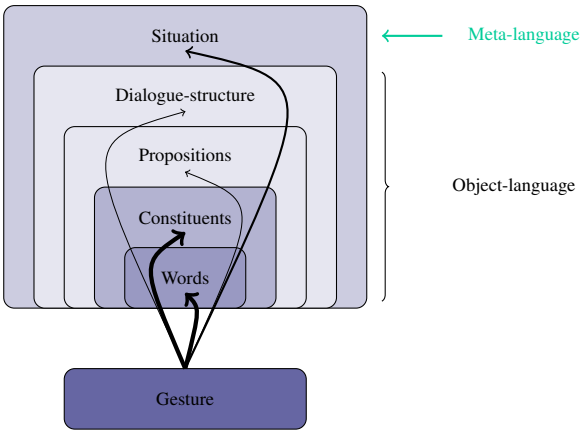


Figure 2: Summary of observations concerning the structural positions and functions of gestures in MM dialogue.[3]

3 A Binding Problem Involving two Representations: How Speech and Gesture Information Are Interfaced

The description of gesture functions provided above may seem fairly convincing, however, we are interested in answering the following questions (a) Do iconic gestures have meaning? (b) Given that they do, how does their meaning interact with verbal meaning? Question (a) has been answered positively in the tradition of semiotic research going back at least to Ch. S. Peirce and carried on in the gesture context by McNeill, Cassell and others. Even if it is difficult to tell how exactly one can provide meanings for gestures on the basis of gesture tokens, we assume here that the representation of gesture mean-

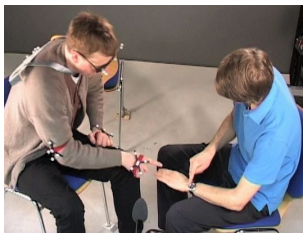
[2]The notion of binding used here is taken from neurobiology and vision research. There is little doubt that the logical notion of operator binding can also be related to these more fundamental notions.

[3]Rev. 1 did not approve of the meta-language label used here. The point is simply that there is no *a priori* argument for putting the formally reconstructed gesture meaning into either the object language or into the specification of the model used. Intuitively, the information of some bi-manual non-symmetric gestures is better placed into a model's definition of domains. One could even start with the hypothesis that gestures generally depict partial models and do not go into the object language at all but investigation of this research line has to wait for another paper.

(a) The concrete sculpture base

(b) Two towers of a church

(c) The U-shape of the town hall



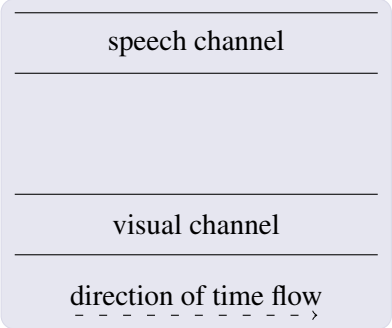
(d) Other correction of follower's gesture

(e) The chapel and a neighbouring tree

Figure 1: Stills showing structural positions of iconic gestures in MM dialogue.

ing can be given in much the same way as for verbal tokens. As a first orientation, assume that gesture meaning behaves functionally like the meaning of deictic expressions. Turning to (b), we will gradually develop a workable schema for a speech-gesture interface below. Starting from the folklore assumption that speech and gesture sit on different channels, we get the picture in Fig. 3, with two channels running in parallel and no interaction specified between speech and gesture. This is meant to serve as our didactic starting point to be modified in stepwise fashion.

Figure 3: Speech channel and visual channel running in parallel



However, there must be an interaction of some sort, since non-lexicalized iconic gestures cannot provide a semantics on their own, so the argument goes in some of the literature (see Kopp et al. (2004) and Lascarides and Stone (2006)). Now the interaction could be of different sorts, e.g. it might be the case that (a1) we can construct some total object language meaning out of the two sorts of meanings or that (a2) we consider one type of meaning as a context to interpret the other type of meaning. An extreme version of (a2) delegates gesture meaning to the context, in particular, to the specification of the model, over which the object language expression is interpreted. So, the function of the MM meaning produced or observed is split, some part goes into the object language and the other into the meta-language. Fig. 2 above indicates that data tell us, when to regard gesture meaning as part of the object language and when to consider it as part of the model. (a2) has as a consequence that one considers only models which satisfy the information provided by the gesture. As a matter of fact, we get most of the information needed for our design decision for an object language (a1) or a meta-language (a2) solution from the annotation depicted in Fig. 4.

[4]The annotation follows two working manuals (Bergmann *et al.* (2007b) for practices and Bergmann *et al.* (2008) for handshapes). The six researchers annotating have been trained over some month on ample raw data; their rate of agreement was

Start Time	End Time
0:39.170	0:41.780
Right.Handshape.Shape	large C
Right.Path.of.Handshape	0
Right.Handshape.Movement.Direction	0
Right.Handshape.Movement.Repetition	0
Right.Palm.Direction	PAB
Right.Path.of.Palm.Direction	0
Right.Palm.Direction.Movement.Direction	0
Right.Palm.Direction.Movement.Repetition	0
Right.Back.of.Hand.Direction	BAB/BUP
Right.Path.of.Back.of.Hand.Direction	0
Right.Back.of.Hand.Direction.Movement.Direction	0
Right.Back.of.Hand.Direction.Movement.Repetition	0
Right.Path.of.Wrist.Location	ARC
Right.Wrist.Location.Movement.Direction	$MR > MF$
Right.Wrist.Location.Movement.Repetition	0
Right.Extent	medium
Right.Temporal.Sequence	0
Left.Handshape.Shape	large C
Left.Path.of.Handshape	0
Left.Handshape.Movement.Direction	0
Left.Handshape.Movement.Repetition	0
Left.Palm.Direction	PAB
Left.Path.of.Palm.Direction	0
Left.Palm.Direction.Movement.Direction	0
Left.Palm.Direction.Movement.Repetition	0
Left.Back.of.Hand.Direction	BAB/BUP
Left.Path.of.Back.of.Hand.Direction	0
Left.Back.of.Hand.Direction.Movement.Direction	0
Left.Back.of.Hand.Direction.Movement.Repetition	0
Left.Path.of.Wrist.Location	ARC
Left.Wrist.Location.Movement.Direction	$ML > MF$
Left.Wrist.Location.Movement.Repetition	0
Left.Extent	medium
Left.Temporal.Sequence	0
Two.Handed.Configuration	$FTT > BHA$
Movement.realtive.to.other.hand	mirror-sagittal

Figure 4: Annotation of example: router's contribution (1) *die Skulptur die die hat 'n Betonsockel / the sculpture it it has a concrete base* [4]

The annotation specifies features and functions of the router's left and right hand, both, on a more global level (the so-called practices like indexing, shaping or grasping giving the global function of the gesture) and on a more fine-grained level which captures the postures of both hands, their parts (palm, back-of-hand, wrist etc.) and their respective movements (left, right, forward etc.). However, the most important thing in the annotation grid is that it maps speech and gesture onto a time line; hence, we can see which speech occurrences overlap with which gesture occurrences. Intuitively, we consider the flowing time as more basic information by help of which speech and gesture events can communicate. Communication among events on different channels is brought about or even caused by temporal synchronization of inputs. This is the concept of binding referred to above. There are several supporting arguments for the binding of gesture meaning to verbal meaning and vice versa:

(1) McNeill (1995, pp. 26-31) considers the stroke information as the carrier of the central semantic and pragmatic information of the gesture. It is in turn tied to the corresponding constituent's stressed syllable or, as we prefer to put it, 'aligned', i.e. synchronized with it. See (Lücking, Rieser, Stegmann (2004)) for experimental evidence.

Supporting arguments (2) and (3) operate on a neuro-information level, (2) concerns vision and (3) cognition in general:

(2) Neuro-biological research on vision is devoted to the so-called binding problem, the dominant model entertained being the time-coding model: the temporal synchronization of the stimuli is the decisive mechanism for integration (Detel (2007), p. 33, translated by the author).

(3) *[Likewise] events that coincide in time are interpreted with greater probability as [being] related than events separated in time* (Singer (1999)).

(4) Finally, from a Gestaltist perspective, rules of grouping and proximity apply.

We cannot enter the difficult problem of neural representation here but will stick to the tools of linguistics and philosophy of language. A rough picture illustrating the information flow of synchronized (aligned) information still using the channel concept is provided in fig. 5. It shows that if there is temporal alignment among events from the different channels, then information from the gesture channel is coordinated with the information from the verbal channel by binding.

4 Interface of Gesture Meaning and Verbal Meaning

We now follow the research strategy a1 introduced in sect. 3. From fig. 5 we see that the following is needed to model the interaction of gesture and

tested in several studies and amounted to 80% in most cases. The temporal boundaries used in example 4 were rated.

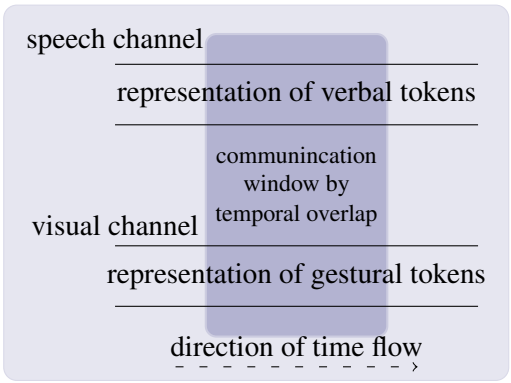


Figure 5: Binding in between the gestural and the verbal channel depending on time synchrony.

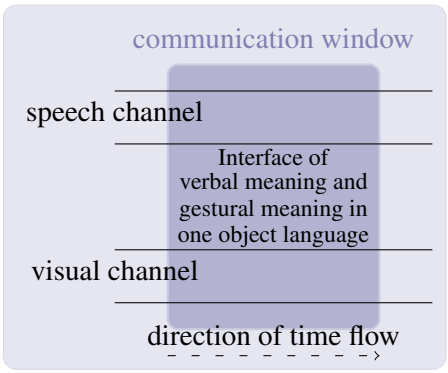


Figure 6: Interface in the communication window established in between the channels.

speech: a representation of (a) the verbal information, of (b) the gesture information compatible with 'Marr structures' (Marr (1982)), and (c) a point of contact for linking the different types of information. (a), (b), and (c) can be achieved using type logics or unification. Gesture information is drawn from the descriptive predicates and values of the fine-grained annotation. For reasons of simplicity we can regard the verbal information as the function operating on the information of the gesture level. However, both must be conceived of as dynamic, due to the direction of time flow on both channels. These inherent constraints can be met by several types of Dynamic Semantics, *inter alia* classical DRT, SDRT, Muskens LDG and PTT, all these add information updating already existing information. The point of contact between the verbal and the gestural level is provided by the window given by temporal overlap (see fig.5), hence temporal synchrony is what matters (i.e. regarded as a necessary condition).[5] The methodological grid now emerging is shown in fig. 6: verbal information and gesture information are interfaced and establish together the context for new information to be integrated. Integration will be anticipated by open slots in the already existing information.

We now specify the procedures for the annotation example in some more detail and concentrate on extracting the semantics of the gesture out of the annotation predicates; the representation of the dynamic semantics of the verbal contribution *die Skulptur die die hat 'n Betonsockel / the sculpture it it has a concrete base* is far from trivial, but we gloss over it here. In the MM example we have the temporal overlap between *Betonsockel/concrete* base and the gesture shown in fig. 1 (a). So, the necessary condition for a fusion of the verbal meaning and the gestural meaning is given, meeting hypotheses (2) - (4) in sec. 3. What do the hands involved sign or inscribe? Here we consider only the relevant parameters in the stroke phase, meeting in particular McNeill's hypothesis (1). The parameters and their values are represented as typed feature structures with types written in italics and standard attribute value pairs <attribute value> used (fig.7).

The matrices show postures of the router's left and right hand as well as two-handed postures. In methodological terms, the annotation predicates constitute the observational language which provides the foundation for our theoretical terms, i.e. the semantic predicates. Figure 8 shows the volume or space shaped by both hands using the annotation predicates as labels.

So, what do both hands depict? Looking at the *R.G.Left* and the *R.G.Right* information, we see that the wrists follow ARC paths. In the beginning, fingers and thumbs touch ($= FTT$), but they separate immediately ($= \neg FTT$). The C-shapes on both hands provide us with a dense series of verticality informations. They also indicate some of the information of a top and a bottom (marked by the top- and bottom-curves of C respectively). $ML > MF$

[5] Rev. 1 does not agree with this assumption, whereas rev. 3 finds it trivial. However, temporal relation of events is the most conspicuous information we can get hold of in the observational data. The ultimate evidence is, of course, a consistent formal model, cf. the remarks in section 2 *A Binding Problem etc*.

$$
\begin{bmatrix}
\textit{Both hands} & & \\
\begin{bmatrix}
\textit{R.G.Left} & \\
\text{HandShape} & \textit{loose C} \\
\text{Palm Direction} & \textit{PAB} \\
\text{BackofHand} & \textit{BAB/BUP} \\
\text{PathofWrist} & \textit{ARC} \\
\text{WristLocation} & \textit{ML} > \textit{MF}
\end{bmatrix} & \begin{bmatrix}
\textit{R.G.Right} & \\
\text{HandShape} & \textit{loose C} \\
\text{Palm Direction} & \textit{PAB} \\
\text{BackofHand} & \textit{BAB/BUP} \\
\text{PathofWrist} & \textit{ARC} \\
\text{WristLocation} & \textit{MR} > \textit{MF}
\end{bmatrix} \\
\text{Two-handedConfiguration} & \textit{FTT} > \neg \textit{FTT} \\
\text{Movement relative to other hand} & \textit{Mirror-sagittal}
\end{bmatrix}
$$

Figure 7: Typed feature structures for some of the information provided in the annotation of fig. 4.

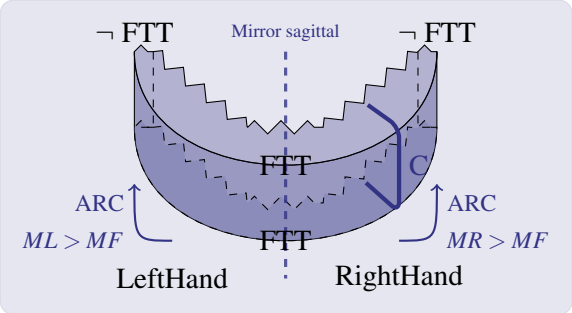


Figure 8: Typed feature structures for some of the information provided in the annotation of fig. 4.

(left forward) and $MR > MF$ (right forward) trace the extent of the curved lines of the sectors bounded by ARC lines. PalmDirection values and BackofHand values follow from the ARC and the WristLocation predicates. We have wrist movements to the left and the right. Finally, Mirror-sagittal shows symmetric extent of the left and the right segment from the router's perspective. What we do now is provide a mapping from the descriptive annotation predicates into a semantic domain. It must specify the depictional value of the gestures and also fix their iconic functions. Thus, the notion of 'similarity' is eliminated *via* a semantic interpretation. Mappings like these have been argued for in (Rieser (2004)) and in (Lascarides and Stone (2006)). We assume a conventional basis for these mappings in Grice's or Lewis' sense, which might depend on a class of contexts: obviously, there must be a reason why we understand gestures and can reliably annotate occurrences of them. The function *rep* indicates representation. *rep* goes from the set of annotation predicates into open formulas. So, the denotation for gestures is provided via translation.[6]

(2) (a) $rep(\text{HandShape } looseC) = hight(x,u) \wedge top(t,u) \wedge bottom(b,u)$

(b) $rep(\text{PathofWrist } ARC) = curved\text{-}side(s,u)$

(c) $rep(\text{WristLocat } ML > MF) = curved\text{-}side\text{-}left(sl,u,router)$

(d) $rep(\text{WristLocat } MR > MF) = curved\text{-}side\text{-}right(sr,u,router)$

(e) $rep(\text{Movement relative to other hand } Mirror\text{-}sagittal) = part(p1,u) \wedge part(p2,u) \wedge (p1 \neq p2) \wedge (p1 \otimes p2) = u$ [7]

In (c) and (d) the routers perspective is coded because of the direction information requiring a *Bühler origo*. The function *rep* induces a mapping from the gesture space GS onto a semantic space SGS.

5 Canonical Word Meaning and How it Can be Extended Using Gesture Information

For purposes of illustration we now assume the following word meaning for *concrete base/Betonsockel*:

(3) $concretebase(x) := support(x,y) \wedge made\text{-}of\text{-}concrete(x) \wedge rigid(x) \wedge object(y) \wedge (x \neq y) \wedge hight(h,x) \wedge side(s,x) \wedge top(t,x) \wedge bottom(b,x).$

[6] (Taking up remarks by all three reviewers). Two problems should be mentioned here. The mapping *rep* is based on observations. It should doubtlessly be backed by statistical data, which are, as yet, not available. Another interesting point is which formal language should be used to represent the gesture meaning. Here I'm still experimenting (cf. also foot-note 7 on fusion). Looking into versions of Mereotopology (see Casati and Varsi (1999) for an overview), I find, that the standard systems available are not strong enough to represent indexical spatial gestures.

[7] The conjunct $(p1 \otimes p2) = u$ is read as 'parts $p1 \otimes p2$ fused yield the whole u', a suggestion I owe to A. Lücking.

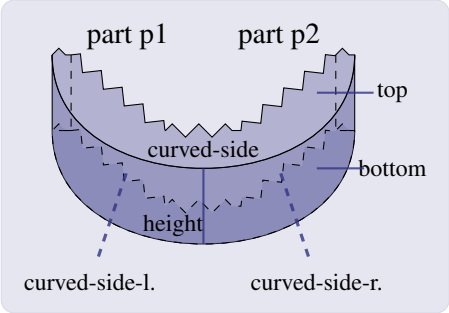


Figure 9: Semantic space SGS induced by gesture space GS, curved lines indicating partiality

So, a concrete base is a support x for an object y iff[8] it is made of concrete, has hight, a side, a top and a bottom. Now, (3) may well be too rich a word meaning for *concrete base/Betonsockel*. So we reduce it and provide an open slot *gest* for the conjunction of the contextual gestural information coded by λ-abstraction in the following way:

(4) $\lambda gest(concretebase(x) := support(x,y) \wedge made\text{-}of\text{-}concrete(x) \wedge object(y) \wedge rigid(x) \wedge (x \neq y) \wedge gest)$

The idea is to model binding between the verbal meaning and the gesture meaning using functional application of (4) for the right-hand-side of (2) as the argument. Hence (4) acts as a context for the gesture information and consumes it. We get

(5) $concretebase(x) := support(x,y) \wedge made\text{-}of\text{-}concrete(x) \wedge object(y) \wedge rigid(x) \wedge (x \neq y) \wedge hight(z,u) \wedge top(t,u) \wedge bottom(b,u) \wedge curved\text{-}side(s,u) \wedge curved\text{-}side\text{-}left(sl,u,router) \wedge curved\text{-}side\text{-}right(sr,u,router) \wedge part(p1,u) \wedge part(p2,u) \wedge (p1 \neq p2) \wedge (p1 \otimes p2) = u.$

What we want to show is:

(a) Contextually, we can do with a minimal word meaning for *concrete base/Betonsockel* consisting of concrete support x for an object y.

(b) Word meaning and gesture meaning interact in context due to temporal binding.

(c) The interface between word meaning and gesture meaning gives us the contextually needed MM meaning which will be more specific than the typical context-free word meaning, and, above all, depend on the situated perspective of the router.

Before we can succeed with (a) - (c), however, we have to deal with the alignment of the objects involved in gesture and speech. Observe that the variables for the logical subjects in (5), x and u, will, as a rule, denote different objects and only contingently refer to the same thing. Intuitively, however, words and gestures in the interface window are about the same object. So, we can formulate the following alignment-of-variables convention:

(6) If words and gestures are about the same object, the same variable must be used for it in the specification of the MM content.

Observing (6) we get an intuitively adequate word meaning.[9]

6 Discussion

In this paper we have only treated the "gesture meaning specifies word meaning" case. We want to take up a few problems. They concern in turn: (1) The reliability of the mapping (2); (2) Dynamic Semantics for lexical information and the embedding of word meaning into the meaning of example (1); (3) Options for reconstructing the relation of word meaning and gestural meaning. Ad (1): Reliability considerations are of course important here, since interpretation and interface construction depend on them. From observation we know that C typically has the function indicated and the same holds for the wrist movements. A slightly different argument in support of the mapping is that one would not find a natural model for example (1) which does not exhibit the gesture semantics indicated. Ad (2): Observe that we can use a dynamic semantics format for the lexical entries. In (5), e.g., we can establish equivalence between two DRSs.We cannot go into matters of establishing a full syntax-semantics interface here, so a few hints must suffice. (7) shows a representation of example (1) in a Muskens LDG format (Muskens

[8]The *iff*-condition will, as a rule, be too strong for word meanings. It is here chosen for reasons of simplicity and perspicuity.

[9](Taking up rev. 1's remarks): All iconic gestures will get a representation using the mapping *rep*. The step from (4) to (5) is computed as described above, modelling *binding between the verbal meaning and the gesture meaning using functional application of (4) for the right-hand-side of (2) as argument*.

(1996)), based on an LTAG representation for reasons of getting at incrementality:

(7) $[x|concretebase(x); \iota y[\ |sculpture(y)] = it; have(it,x)].$

Sticking to the format of explicit definition, we can substitute the right side of expression (5) suitably represented for *concrete base*(x). Hence, intuitively, we will get suitable derivation- and entailment-relations. Ad (3): You may have noticed that the word meaning in (3) does not fully specify the shape of the figure's tops and bottoms. Assume, we add *elliptical(t)* and *elliptical(b)* in order to provide the missing information. Then we run into a problem with (5), since (5) only partially provides the information of an extended (3). It turns out that we encounter a Gestalt regularity here, the principle of Prägnanz (*minimum principle*) being at stake. Perhaps we could solve cases like this one using abduction but it is not trivial to do this. Another Gestalt issue is that gestural movements are not precise in the geometry sense. We leave these topics for a methodology paper.

Acknowledgements

The work reported in this paper has been supported by the German Research Foundation (Project B1, *Speech-Gesture Alignment*, CRC *Alignment in Communication*, Bielefeld University) which is gratefully acknowledged. Thanks go to my co-workers Kirsten Bergmann, Andy Lücking, Florian Hahn and Stefan Kopp for discussion and support.

References

Bergmann, K. and Rieser, H. 2007. Discussion of A. Lascarides and M. Stone's Example (1) from their *Formal Semantics for Iconic Gesture.* Workshop-contribution, Bielefeld Univ., June 2007

Bergmann, K., Fröhlich, C., Hahn, F., Kopp, St., Lücking, A. and Rieser, H. 2007. Wegbeschreibungsexperiment: *Grobannotationsschema.* Bielefeld Univ., June 2007

Bergmann, K., Damm, O., Fröhlich, Hahn, F., Kopp, St., Lücking, A., Rieser, H. and Thomas, N. 2008. *Annotationsmanual zur Gestenmorphologie* Bielefeld Univ., June 2008

Casati, R. and Varsi, A. C. 1999. *Parts and Places. The Structures of Spatial Representation.* The MIT Press: Cambr., Mass.

Detel, W. 2007. *Grundkurs Philosophie, Bd. 4, Erkenntnis- und Wissenschaftstheorie.* Reclam: Stuttgart.

Kita, S. (ed.) 2003. *Pointing. Where Language, Culture, and Cognition Meet.* Erlbaum: London.

Kopp, St., Bergmann, K. and Wachsmuth, I. 2008. *Multimodal Communication From Multimodal Thinking – Towards an Integrated Model of Speech and Gesture Production* In *International Journal of Semantic Computing* (in print).

Kopp, St., Tepper, P. and Cassell, J. 2004. *Towards integrated micro-planning of language and iconic gesture for multimodal output.* In *Proceedings of ICMI.*

Lascarides, A. and Stone, M. 2006. *Formal Semantics for Iconic Gesture.* In *Proceedings of Brandial*. Potsdam University.

Lücking A., Rieser, H. and Stegmann, J. 2004. *Statistical support for the study of structures in multimodal dialogue.* In *Proceedings of Catalog* 04, pp. 56-64

Lücking, A., Pfeiffer, Th. and Rieser, H. 2008. *Pointing Reconsidered.* Submitted

Marr, D. 1982. *Vision.* San Francisco: Freeman

McNeill, D. 1995. *Hand and Mind. What Gestures Reveal about Thought.* UCP: Chicago and London.

McNeill, D. (ed.). 2000. Language and gesture. CUP.

Muskens, G. 1996. *Combining Montague Semantics and Discourse Representation.* In: *Linguistics and Philosophy*, 19, pp. 143-186.

Rieser, H. 2004. *Pointing in Dialogue.* In *Proceedings of Catalog* 04, pp. 93-101

Rieser, H. 2005. *Pointing and Grasping in Concert. With an Encore on Salience.* In: Stede et al. (eds.), *Salience in Discourse: Multidisciplinary Approaches to Discourse* 2005. pp. 129-139

Rieser, H. 2007. *Multimodal action: Demonstration and reference.* In: *IPA Abstracts.* Göteborg, Sweden, pp. 148-149

Rieser, H., Kopp, St. and Wachsmuth, I. 2007. *Speech-Gesture Alignment.* In: *ISGS Abstracts, Integrating Gestures.* Northwestern University, Evanston, Chicago, pp. 25-27

Rieser, H. and Staudacher, M. 2008. *SDRT and Multimodal Situated Communication.* Submitted

Singer, W. 1999. *Neural Synchrony: A Versatile Code for the Definition of Relations.* In *Neuron*, Vol. 24, pp. 49-65.

Discourse Motivated Constraint Prioritisation For Task-Oriented Multi-Party Dialogue Systems

Petra-Maria Strauß, Simon Friedmann, Tobias Heinroth
Institute of Information Technology
University of Ulm
Germany
{petra-maria.strauss;tobias.heinroth}@uni-ulm.de

Abstract

This paper presents a new algorithm to prioritise user constraints for problem solving in task-oriented multi-party dialogues. The situation of (at least) two users pursuing a common goal supersede the need for exhaustive semantic analysis which is commonly used in dialogue systems to prioritise user constraints. Instead, we suggest to use the ongoing discourse, especially the order of occurrence of the constraints for prioritisation. In this paper, we describe our algorithm and the scenario in which it was applied. We further present a first evaluation in which we compared our approach to semantic prioritisation which showed very promising results. It proofed that for our domain our simple algorithm outperformed its opponent by far.

1 Introduction

In this paper we present a new algorithm to prioritise user constraints for problem solving in task-oriented multi-party dialogue systems. Most prevailing spoken language dialogue systems (SLDS) are single-user systems. One user is interacting with the computer while the computer collects the information provided by the user. Multi-party systems pose new challenges, however, also new opportunities considering their characteristics. The multi-party SLDS interacts with more than one user, the users interact with the computer and additionally with each other. Thus, the system not only needs to understand the requests posed directly towards it but additionally has to follow the dialogue between the users in order to comprehend the entire conversation and grasp the context.

Naturally, the conversation partners often have different preferences which complicate the problem solving process immensely. The course of the dialogue also depends on the sort of dialogue and domain. In our example domain of restaurant selection two human dialogue partners and a computer interact with each other. The dialogue partners are generally not interested in a long discussion but rather in coming to a quick consensus. Besides uttering their own preferences and dislikes, the dialogue partners evaluate each other's preferences against their own and react accordingly.

The system therefore does not model the preferences of each user independently but collects all information relevant for the task to form a set of so-called constraints for the data base queries. If the query does not yield any results ('over-constraint situation'), an intelligent system is expected to provide an alternative solution. The common approach to that is to relax less important constraints. In single-user systems the prioritising of constraints is mainly performed by semantically analysing the constraint-bearing utterances in terms of keywords that denote the importance of the constraint to the user. However, the collection of valid constraints is also more straight-forward, depending on the preferences of the single user and on data-base constraints.

In the multi-party case, in the course of the dialogue each introduced constraint is discussed, rejected or accepted by the other dialogue partner. This makes automatic semantic analysis very complicated. We claim that in this case the prioritisation

process can be a lot simpler. We take the content of the discourse into account, i.e. the longer a constraint is valid in the dialogue, the more important it gets.

Various research groups have been working on the same domain of restaurant selection. For the system to cooperatively find a suitable restaurant, it has been found to be of utmost importance to consider the users' preferences, as well as also the strengths of these preferences (e.g. (Carberry et al., 1999)). Work on the closely related matter of presenting information and options in a SLDS can be found e.g. in (Demberg and Moore, 2006), (Walker et al., 2004), and (Carenini and Moore, 2001).

However, all of these surveyed dialogue systems are single-user systems. We state in the following sections why and how the multi-party situation differs from prevalent single-user systems and why it gives rise to new approaches. We briefly present the scenario and dialogue system in which we deploy the presented approach in Section 2. Section 3 focuses on different aspects of multi-party interaction. Section 4 introduces our approach to prioritisation exploiting multi-party characteristics to enhance the constraint based problem solving used in our system. The evaluation is described in Section 5 before Section 6 concludes the paper.

2 Multi-Party Dialogue System

Two human dialogue partners interact with a computer which acts as an independent dialogue partner in the scenario of restaurant selection (Strauss, 2006). In the beginning of the dialogue, the users talk about an optional topic while the system passively observes and captures the relevant conversational context. As soon as the users come to speak of the specified domain the system starts to "listen" attentively. When required by the conversational situation, it takes the initiative to get meaningfully involved in the communication and to help solve the task, i.e. help the users to find a suitable restaurant.

The analyses presented in this paper were performed on a set of dialogues from a corpus obtained through Wizard-of-Oz recordings (Strauss et al., 2008). The dialogues were transcribed and annotated with a simple tagset of 10 dialogue acts (refer to Section 3.2).

3 Multi-Party Interaction

Multi-party dialogue systems differ in many ways from single-user systems. The counterparts of the system are at least two people who interact with the system and additionally with each other. The face-to-face interaction with human dialogue partners is for humans still the most natural and comfortable way of communicating. Presumably, it is also faster and more efficient, e.g. due to the human ability of dissolving ambiguity by interpreting paralinguistic phenomena of communication such as emotions and facial expressions of the other dialogue partner. Thus, multi-party dialogue systems can be very advantageous in terms of that humans still communicate with each other and additionally turn towards the system only when necessary. The system is acting only as a side-participant of the conversation when the users don't need it.

Consequently, the design of our system is convenient as the users are able to first come to an initial agreement among themselves before the system gets involved in the conversation. This seems more efficient and faster than if they would have been interacting with the system during the entire process.

A further point crucial for the system's usability is the process of problem solving itself and how the results are presented to the users. Consideration and prioritisation of the users' preferences is an important issue in this respect. The research addressing this problem has so far been only considering single-user situation (e.g. (Carberry et al., 1999)). Before we present a new approach for the multi-party situation that utilises all the benefits that come with the additional dialogue partner, we discuss a few more points relevant for multi-party interaction.

3.1 Communication Roles

The situation in which the communication takes place has a substantial impact on the conversation. Next to the conversational roles such as speaker, addressee and overhearer (Clark, 1996), the roles the participants take on socially in the conversation play an important role in dialogues. (Bunt, 1994) introduced the social context as part of the dialogue context. (Traum, 2004) talks about the specific task roles which relate dialogue participants in certain ways.

Utterance	DA, Reference
A5: Let's go to an Italian restaurant.	(sugg,{})
B6: An Italian restaurant?	(check,A5)
A7: Yes.	(ack,B6)
B8: Ok.	(acc,A5)

Table 1: Dialogue snippet

During the Wizard-of-Oz recordings we randomly assigned different roles and scenarios to the dialogue partners. These included e.g. employer and employee, lovers, business colleagues, or friends. This way, we tried to obtain a wide range of different (such as superior / inferior) behaviour in our corpus which is important to be able to evaluate our approach on a broad variety of dialogues.

3.2 Interaction Phenomena

Task-oriented human-human dialogue shows a certain pattern which needs to be understood by the system to be able to model the conversation. The users mainly exchange proposals, introducing their preferences into the conversation. A proposal from one of the dialogue partners induces a reaction from the other dialogue participant. This response may consist of a simple acknowledgement, an accept or reject, a response with further content, or possibly a counter-proposal. Sometimes, the dialogue partner repeats the proposal which can have the function of acknowledgement, of checking if it was understood correctly or as a way of deferring the dialogue in order to win time to think. The response does not necessarily follow up a proposal but can also occur various turns later in the conversation with possibly even talking about a different topic in the meantime.

Table 1 shows a short example dialogue snippet labelled with the according dialogue act and the number of the utterance it refers to. We deploy a tagset of 10 basic dialogue acts which satisfies our domain and dialogue system requirements: *request, suggest, inform, acknowledge, check, accept, reject, stall, greet, other*.

User A proposes to go to an Italian restaurant. Instead of accepting right away, *User B* repeats *A*'s proposal whereupon *A* acknowledges *B*'s repetition. In this case, the repetition is to be interpreted as a request for clarification (check act).

4 Discourse Motivated Constraint Prioritisation

During the course of the conversation, the system collects all information relevant for the task which forms the basis for the database query and thus narrows down the result set in terms of positive or negative constraints.

If no results are obtained, i.e. an over-constraint situation occurred, the system should offer the users an alternative result. For this, we deploy constraint prioritisation in order to take user preferences into account. Different approaches to user preferences have been introduced (e.g. (Carberry et al., 1999)), however, only for single user dialogue systems. In a single-user system, finding out user preferences can be done using different methods. One is to analyse the semantic content of an utterance looking for specific words that show some kind of sign of importance, e.g. 'maybe', 'definitely', etc.[1] Another way is to simply ask the user about which constraint is more important in case that the system encounters an over-constraint situation.

In contrast, a multi-party system has one big advantage over all single-user systems: The additional - human - dialogue partner who already analyses the utterances from the other dialogue partner. When a suggestion is introduced with a 'maybe', this low priority is recognised by the dialogue partner, who can then accept the suggestion, or, being aware of the low priority of the proposal, make his or her own counter-suggestion which might be more precise.

A request to the computer is generally expressed only when the dialogue participants have found their highest common priorities. In the following, we describe the algorithm in detail.

4.1 Prioritisation Scheme

For prioritisation, information is extracted of each utterances according to three categories: Changing Categories, Current Preferences, and Prioritisation Values.

[1] Actually, it is not that simple as e.g. 'inferior' words are also more likely uttered by 'inferior' dialogue participants. However, a further elaboration on this aspect goes beyond the scope of this paper.

Changing Categories (CC) indicate the topic(s) of the current utterance. For instance, if one of the participants makes the statement of wanting to eat Italian food, the CC field is tagged with category (**F**) which stands for food or cuisine.

The other distinguishable categories are location (**L**), ambiance (**A**), category (**C**), price range (**P**), specials (**S**) and opening hours (**O**).

Current Preferences (CP) lists all currently valid constraints represented by individuals of the respective category and is thus used for a database query. In the example above, 'Italian' would be categorised as food (F) and individual **F1** (taken it is the first F-subject in this conversation). A second F-value later on in the dialogue, e.g. Mexican food, would then be tagged **F2**, etc. This is applied analogously to all other categories (L1, L2, P1 etc.).

Prioritisation Values (PV) assign a priority value to every individual. With every recalculation (induced by a change in the CP section) all currently valid values rise by '1 point'. A new individual is introduced with the value '1', i.e. it has risen '1 point' from the default value of '0'. Negative constraints or dislikes are represented with negative values accordingly (starting at '-1').

4.2 Executing the Prioritisation Scheme

In the following, the prioritisation algorithm is applied to a dialogue.

Introducing Preferences

At the beginning of a dialogue, the table contains no entries. As soon as a topic is raised, it is displayed in the CC section. The corresponding individual is inserted into the CP column of the table and the PV value is '1' (or '-1' in case of negation).

During the Dialogue

Every time the users modify their constraints, e.g. by proposing or dismissing one, a change in the CP section occurs and the PV are recalculated: The values of all individuals that are currently represented as valid preferences (in CP) are raised by '1' (or lowered by '1' for negative values).

Thus, the longer a subject stays valid, the higher its priority value becomes, which is obviously the desired effect. That means, as long as a subject is not explicitly abandoned or replaced by a different value due to incompatibility between constraints, it is considered valid and part of the current preferences. If a constraint is dismissed it is taken out of CP, its PV stays at the current value. Should it be re-introduced in the dialogue with the same polarity, it is reinserted into CP and the priority calculation starts at the former value. A change in the polarity of a valid constraint is performed by simply adding or removing the '-'.

System Involvement

All currently valid individuals are listed in the CP section which serves as the basis for the system's database queries. Every change in the constraint set induces a database query so the system is always up-to-date and ready to interact. Generally, the system interacts for the first time after the users have already come to an initial agreement. As also noted by (Carberry et al., 1999), this first request to the computer deserves special attention as it displays the users' original preference. Thus, all valid individuals at the time of the first computer request receive a *first request bonus* of '2'. This number provides an adequate trade-off between raising the priorities enough to stand out, but at the same time not too high so they can still be 'overruled', if necessary.

Table 2 shows a short part of a dialogue. At the beginning of this dialogue snippet, new individuals are introduced, namely 'Spanish' and 'Italian', 'F2' and 'F3'. It can be seen that the highest priority for the users at that point have the location 'L1' and the ambiance 'A1'. Both are at value '5', which means they assumably were the first constraints to be introduced in the dialogue and also received 2 points bonus. The second displayed utterance does not induce a recalculation of new preferences, due to the fact that nothing has changed. The request is simply repeated by the other dialogue participant. In the following utterance, the introduction of German cuisine as a negative value induces a change in CP and triggers a recalculation of the priorities.

4.3 Applying Prioritisation to Constraint Based Problem Solving

At present, the prioritisation only comes into play in the case of an over-constraint situation, i.e. if the database query does not yield any results. In order to offer the users a best possible alternative result

Utterance	CC	CP	PV	
...				
B: I would like something Italian or Spanish.	F	F2 F3 C1 L1 A1	F2 = 1 F3 = 1 C1 = 4 L1 = 5 A1 = 5	(italian) (spanish) (restaurant) (river) (beergarden)
A: Italian or Spanish is fine with me.				
A: I just don't want German food.	F	F4 F2 F3 C1 L1 A1	F4 =-1 F2 = 2 F3 = 2 C1 = 5 L1 = 6 A1 = 6	(german) (italian) (spanish) (restaurant) (river) (beergarden)
...				

Table 2: Prioritisation scheme applied to an extract of a dialogue.

the system has to decide which constraint(s) to relax. We deploy the following (simplified) algorithm:

```
while overconstraint OR resultset ==
previous_resultset do
    if onto_check(relax_candidate).succeed then
        present_results();
        break;
    else
        if relax(relax_candidate).succeed then
            present_results();
            break;
        end
    end
    relax_candidate++;
end
```

Algorithm 1: Simplified relaxation algorithm

The constraint with the lowest priority value is chosen as the first relaxation candidate. The result of the following query is analysed in terms of another over-constraint situation. The result set is further compared to the result set that was presented to the users in the system's last turn before the initial over-constraint situation occurred. If the result sets are the same, i.e. the same result set that obviously had just been rejected or further constrained by the users would be presented again. Thus, the relaxation algorithm proceeds at this point. If again no result was obtained, the relaxed value is reinserted before the next relaxing candidate is considered for relaxation. After another unsatisfying result, both values are relaxed etc. The presented algorithm is simplified in this matter and also in the way that it assumes that each time there is exactly one constraint with minimal priority value which, however, is not always the case. The implemented algorithm handles this by trying out each of the potential relaxation candidates and taking the one with the best results.

Before relaxing, the relaxation candidate is inspected in the context of the ontology to take related values into account. If e.g. no restaurant can be found near the town-hall before relaxing this constraint, it will be checked if there would possibly be one around the cathedral which is the adjacent area. This kind of ontology check can be performed for all exclusive categories (L, F, P, C, and A). However, the observation of the recorded dialogues showed that some values should not be relaxed if possible. They include e.g. the values of category *S* (i.e. 'specials', such as cocktails), or 'expensive' of category *P*, as well as negative constraints. No matter at what point these values were introduced in the dialogue, they were very important to the users and therefore not relaxed.

5 Evaluation

We performed evaluation on a set of dialogues from our corpus (Strauss et al., 2008). In the normal course of a dialogue all constraints are considered in each database query regardless of their priority. Therefore, evaluation can only be performed on di-

alogues where over-constraint situations occurred. This resulted in a set of 14 dialogues.

At recording time, the system simply told the users that there were no results found. The users then modified their query according to their preferences. For evaluation we compared the outcome of our algorithm to the users' reaction to the over-constraint situation and how they proceeded in the dialogue, i.e. which constraints they relaxed or modified. The relaxation algorithm performed equally well or better in 13 out of 14 dialogues, i.e. the algorithm leaded to relaxing the same constraints as the users did. By conducting the ontology check, in 5 of the 13 cases the outcome would have even been better as the system would have suggested a result closer to the original preferences than what was obtained in the dialogue.

We further compared the performance of our algorithm to semantic prioritisation. For this, we hand-annotated the constraints (mainly by considering keywords that denote importance) using a weighting scheme from '1' (little interest) to '5' (strongest interest). The same range is applied to dislikes ('-5' to '-1', with '-5' meaning strongest dislike). Weights were dynamically adapted during the course of the dialogue, if necessary. The semantic algorithm performed as well as ours in 6 cases. In most cases it relaxed the constraints in a different order which mostly also lead to a different result set. The semantic algorithm repeatedly tried to relax one or more of the users' main preferences which e.g. becomes apparent in one of the dialogues just after the over-constraint situation when the users tried to rephrase their main preferences which at this point the system already would have had relaxed. In 1 case, the semantic algorithm performed better than ours in the way that it relaxed the same constraint as the users when ours did not.

The overall result is therefore very affirmative: Our algorithm represents user preferences equally well or better than a similar method using semantic analysis for prioritising user constraints in all but one evaluated cases.

6 Conclusion

In this paper we presented a new algorithm to prioritise user preferences in a task-oriented multi-party dialogue system. We use the ongoing dialogue to assign priority values to the constraints, i.e. the longer a constraint is valid in the dialogue the more important it gets. The evaluation of our simple approach showed auspicious performance. We compared it to a semantic prioritisation approach as well as to how the users actually proceeded after an over-constraint situation had occurred in the analysed dialogues. In all but one case, our algorithm performed equally well or better.

Future work includes further evaluation, also using different domains. Additionally, we are planning to take the frequency of changes in a certain category into account. For instance, if the users switch many times between different kinds of cuisine, the value for this category would be rather high and imply a sort of uncertainty and flexibility in this aspect.

References

Harry C. Bunt. 1994. *Context and Dialogue Control*. THINK Quarterly, vol.3, pp.19-31.

S. Carberry and J. Chu-Carroll and S. Elzer. 1999. *Constructing and Utilizing a Model of User Preferences in Collaborative Consultation Dialogues*. Journal of Computational Intelligence, vol.15, no.3, pp.185-217.

G. Carenini and J.D. Moore. 2001. *An Empirical Study of the Influence of User Tailoring on Evaluative Argument Effectiveness*. IJCAI, pp.1307-1314. Seattle, Washington, USA.

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

V. Demberg and J.D. Moore. 2006. *Information Presentation in Spoken Dialogue Systems*. EACL, pp.65-72. Trento, Italy.

P.-M. Strauss. 2006. *A SLDS for Perception and Interaction in Multi-User Environments*. 2nd International Conference on Intelligent Environments (IE06). Athens, Greece.

P.-M. Strauss and H. Hoffmann and W. Minker and H. Neumann and G. Palm and S. Scherer and H. C. Traue and U. Weidenbacher. 2008. *The PIT Corpus Of German Multi-Party Dialogues*. International Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco.

D. Traum. 2004. *Issues in Multiparty Dialogues*. Advances in Agent Communication Ed. F. Dignum. Springer-Verlag LNAI 2922, pp.201-211.

M.A. Walker and S.J. Whittaker and A. Stent and P. Maloor and J.D. Moore and M. Johnston and G. Vasireddy. 2004. *Generation and evaluation of user-tailored responses in multimodal dialogue*. Journal of Cognitive Science, vol.28, pp.811-840.

Resolving Ambiguous, Implicit and Non-Literal References by Jointly Reasoning over Linguistic and Non-Linguistic Knowledge

Nicholas L. Cassimatis
Department of Cognitive Science, Rensselaer Polytechnic Institute
Troy, NY 12180 USA
cassin@rpi.edu

Abstract

The problem of resolving ambiguous, implicit and non-literal references exemplifies many difficult issues in understanding language. We describe an approach for dealing with these by representing and jointly reasoning over linguistic and non-linguistic knowledge (including structures such as scripts and frames) within the same inference framework. This approach enables a treatment of several reference resolution phenomena that to our knowledge have not previously been the subject of a unified analysis. These results suggest that treating language understanding as an inference problem encompassing nonlinguistic knowledge can expand the ability of computational systems to use language.

1 Difficult references

Ambiguous, implicit and non-literal references embody several difficult problems in language understanding. We propose an approach for dealing with these problems that involves representing and jointly reasoning over syntactic, semantic and nonlinguistic knowledge.

Formal and computational accounts of language use have difficulties with utterances whose meaning cannot be compactly and unambiguously captured as a function the meanings of the elements of an utterance. Many of these problems are evident in reference resolution and they involve the interaction of linguistic and nonlinguistic information. The following examples (adapted from (Hobbs, Stickel, Appelt, & Martin, 1990)) illustrate this:

(1) Dave hid Paul's keys. He was drunk.
(2) Dave hid Paul's keys. He often jokes with him.

Finding the most likely antecedent for "he" in each sentence depends in part on nonlinguistic factors such as the relationship of the people discussed, the necessity of keys in driving and the effects of drunken driving.

The following cases (copied or adapted from (McShane, in preparation)) illustrate references items not explicitly mentioned in discourse.

(3) The couple went for a walk. He held her hand. (Referent of "he" is member of set).
(4) The home goalie played hard but the visiting goalie played even harder. Both of them got special mention after the game. (Referent of "both" is set formed by previously mentioned objects).
(5) The storm lasted for hours. The thunder scared my dog. (Referent for "thunder" implied by noun).
(6) It thundered for hours. The thunder scares my dog. (Verbal antecedent).
(7) George Bush signed the bill. (Referent of "George Bush" is in common knowledge).

Finally, the actual reference of an utterance can be entirely different from the literal reference:

(8) The author began the book. (Pustejovsky, 1995) (The *writing* of the book was begun.)
(9) The ham sandwich ordered some coffee. (Nunberg, 1979) (The person who is eating the ham sandwich ordered the coffee).

Several approaches have been used to resolve references computationally. One approach is first the "structured knowledge" approach. Relatively large and complex structures such as scripts (Schank & Abelson, 1977) and frames (Minsky, 1975) can be used to encode much of the knowledge needed to resolve references. For example, consider a typical frame-based account of (3). It

presumes a "couple frame" that has two slots, one for each member. One slot (M) is marked as male and the other (F) as female. During processing, when "couple" is encountered, an instance of the couple frame is instantiated. When "he" and "she" are processed, the task is to "match" them to slots in frames that have already been instantiated. In this case, "he" and "she" match the male and female member of the couple frame respectively.

Although capable of dealing with many otherwise difficult cases, the structured knowledge approach has several problems. First, structures often do not work in cases that vary slightly from those for which they were designed. Second, the matching process is not always smooth. For example, in "The couple went for a walk with their daughter. They held her hand", "her" could match the female member of the couple and also the daughter. Matching algorithms that deal with such ambiguities are very complex and imperfect. They cannot easily incorporate "common sense" reasoning that would in this case infer that "her" refers to the daughter since otherwise the couple would be holding the hand of the female member of the couple, which is highly unusual.

The statistical, corpus driven approach to reference resolution (e.g., (Mitkov, 2000)) relies on the premise that there is enough information latent in actual instances of language use to successfully resolves inferences. By not involving complex structures or matching algorithms, they do not raise many of the difficulties of the structured knowledge approach. On some corpora, they can achieve upwards of 90% accuracy. However, many of these results rely on very specific assumptions, e.g., that the antecedent to a referent is explicitly mentioned in the text. However, as (3)-(7) illustrate, there are many cases where antecedents do not occur anywhere in the text. Current corpora cannot be used to deal with such cases because they only mark antecedents that explicitly occur. Additionally, there are many cases where even infants can find referents of novel words for which they have no statistical information. Finally, performance even in cases where corpora can be used has plateaued in the field, suggestion limits to the potential purely statistical approaches.

The "inferential" approach to reference resolution (e.g., (Hobbs et al., 1990)) eases the combination of reasoning over world and linguistic knowledge that seems to be required for reference resolution. The inference approach views utterances as actions taken by people and the problem of language understanding as a kind of action understanding or abduction problem. By formulating both the linguistic and nonlinguistic constraints using the same inferential framework, the hope is that the right meaning for an utterance can be inferred using general-purpose and flexible inference engines rather than precarious structure-matching algorithms. In (1), for example, an abduction process would explain Dave's hiding of Paul's keys by Dave's desire to prevent Paul from inuring himself while driving drunk. Once this has been inferred, then the subject of "was drunk" must refer to Paul and thus the coreference of "he" and "Paul" is inferred.

Although the inferential approach has achieved some success, it has suffered from the lack of powerful enough inference mechanisms has not so far yielded an analysis of non-literal uses of language.

2 An inferential approach

The work described in this paper is based on a new incarnation of the inference approach that is intended to address some of its past deficiencies and broaden the range of linguistic phenomena explainable within a single formal or computational framework. It is based on several precepts:

Action understanding. We adopt the view (Clark, 1996) that conceives of utterances as actions taken by a user and the problem of language understanding as one of finding the best explanation of these actions.

Non-linguistic constraints. We believe that nonlinguistic knowledge and information is often key to inferring the meaning of an utterance. This can include knowledge about the world, people's beliefs and desires and perceptual salience. One consequence is that nonlinguistic knowledge must be part of explaining many linguistic phenomena.

Single inferential substrate. In order to explain how linguistic and nonlinguistic knowledge constrain language understanding, we use the same "substrate" of relations and inference methods to encode and reason over this knowledge. Some constraints "span" linguistic and nonlinguistic information. For example, lexical entries often include phonological information about a word as well as what aspects of the world the word normally refers to. Thus, by combining linguistic, non-

linguistic and spanning constraints into one set of constraints, a reasoning engine that operates over this set will automatically and without any special provision use linguistic and non-linguistic information to constraint interpretation.

Structures as sets of constraints. Although we take the inferential approach, we presume that scripts, frames and other elements from structured knowledge approaches are required to explain language use. We thus encode this knowledge using the same constraint language use to encode other knowledge. Since structures tend to have exceptions (e.g., although a room script would include slots for windows, not all rooms have windows), it is important to use a framework that uses "soft" constraints that can be violated.

3 Inferential framework

Our approach relies on a language for expressing probabilistic constraints over relations among objects. Many aspects of the language has characteristics common to typical logical and probabilistic reasoning frameworks. Although work with such languages typically involves logical or probabilistic reasoning methods (such as MCMC, SAT solving or resolution), we remain agnostic in this paper as to which mechanisms are used. We do suspect however, that analogical, case-based and neural-network methods not normally associated with logical and probabilistic inference will also be required for the kind of inferences we describe here to be made in any kind of realistic scenario.

In this language, constraints are probabilistic conditionals whose antecedent and consequents are possibly negated first-order literals. Variables all start with "?". For example, the constraint, $Wet(?x) + Iron(?x) \rightarrow .95\ Rust(?x)$ states that if something is wet and iron, it has a 95% of rusting because of this. Facts can be stated as constraints with antecedents that are always true, e.g., $True() \rightarrow 1\ Rises(sun)$. This can be abbreviated simple as $Rises(sun)$.

Constraints can also be followed by "posited variables" that license the positing of objects. Consider, for example:

$$Plane(?p) \wedge InRange(?p, ?r) \rightarrow .87\ Blip(?r), ?p$$

This constraint states that a plane in range of a radar station has an 87% chance of causing a blip. In the case where a particular radar station has a blip, one can infer the existence of a plane that caused that blip, even if the plane was not known about in advance.

Finally, we presuppose the ability to find the most likely world(s) given a set of constraints. Specifically, given a set of constraints C, there are several worlds consistent with it. A world is simply an assignment of truth values to the propositions in or licensed by C. For example, if $C = (True(Rain(today)) \rightarrow .8\ \neg Rain(today))$, there are two worlds consistent with that: $w_1 = (Rain(today),\ true)$ and $w_2 = (Rain(today),\ false)$. Worlds have a probability of being actual. The probability of w_1 is .2 and of w_2 is .8.

Finally, identity is an important relation in what follows. $Same(x, y)$ states that x and y name the same object. We will presume the axioms of identity. E.g., if $P(x)$ and $Same(x, y)$, then $P(y)$.

Although apparently straightforward, inference approaches for languages with identity that license the positing of objects raise several difficult technical issues that have not begun to be dealt with until recently (e.g., (Milch et al., 2005)).

4 Fundamentals

Our overall goal is to represent and jointly reason over linguistic and nonlinguistic knowledge in order to provide a unified account of some difficult aspects of language use. This section presents some basic precepts of how to use the inferential framework described in the last section to accomplish this.

4.1 Linguistic knowledge

We will assume that the totality of a language understander's linguistic and non-linguistic knowledge is encoded in a set of constraints, C. In what follows, we illustrate the kinds of constraints our approach uses.

Utterances can be represented using logical propositions. For example, we indicate that "Mary likes John" was uttered with the following propositions:

$$IsA(w1, Word), Phonology(w1, "mary"), Occurs(w1, t1), IsA(w2, Word), \ldots .$$

In our analyses, we presume that the syntactic structure of utterances as given. However, work casting parsing as an inference problem (Murugesan & Cassimatis, 2006) makes us optimistic that syntactic parsing can also be dealt with as an inference problem.

The literal reference and semantic information of a word or phrase can be indicated thus:

$$LitRef(w1, litRef) \text{ ^ } Name(litRef, \text{“}Mary\text{”}) \text{ ^ } IsA(litRef, Female)$$

The literal reference of a word is not always its actual reference. For example, in the case where the speaker means that Mary's dog likes John's dog, we can say: $Ref(w1, dog12)$.

How this reference is determined will be discussed in the next section.

We can represent that that the literal reference is often the actual references with:

$$LitRef(?w, ?litRef) \rightarrow p_{lit}\ Ref(?w, ?litRef).$$

How the actual value of p_{lit}, the probability that phrases are used literally, is arrived at is left for future research. All we assume in what follows that it is relatively close to 1.

Coreference in this framework is an identity relationship. For example, if in "John likes himself" the actual referent of John is *j-ref* and "himself" is *h-ref*, then "John" and "himself" corefer if *Same(h-ref,j-ref)*.

Ambiguity of reference is uncertainty about identity. For example, in (10), the reference of "he" can be John or Fred.

(10) John and Fred are friends because he is rich.

This is represented thus: $Ref(w1, john)$, $Ref(w3, fred), Ref(w7, h)$. "he" can refer to John: $Same(h, john)$ or to $Same(h, fred)$. If we assume the background knowledge that $\neg Same(fred, john)$, then h cannot equal both $fred$ and $john$.

We have thus far presumed several components of C, the constraints representing the listener's knowledge. To summarize, these include the set of utterances (U) heard, syntactic knowledge (SYN), semantic knowledge (SEM) (e.g, that literal reference of Mary is a female named Mary), pragmatic ($PRAG$, e.g., that the literal reference tends to be the actual referent) and non-linguistic knowledge ($WORLD$, e.g., that John is not the same person as Fred).

The debate as to whether or how to precisely distinguish between syntactic, semantic and pragmatic knowledge need not be settled to proceed with our analyses. All forms of knowledge, linguistic and non-linguistic, are treated identically within the inferential framework we are using. They are all simply constraints.

4.2 Non-linguistic knowledge

How to represent the full range of non-linguistic knowledge is of course a very broad and difficult problem. However, for our purposes, it is enough to describe how we use constraints to represent structures such as frames and scripts.

Frames and scripts can both be characterized in terms of "slot-filler" pairs together with properties of the fillers and relationships between the fillers. For example, imagine a couple frame that has two slots. The filler of one slot has the property of being a male and the filler of the other is a female.

The information in scripts and frames can be captured by constraints. For example, the following constraints represent the information encoded in the couple frame:

$$Couple(?c) \rightarrow 1$$
$$PartOf(?m, ?c) \text{ ^ } PartOf(?m, ?c), ?m, ?f$$

(Couple frames have two slots, each of whose filler is a part of the couple).

$$Couple(?c) \text{ ^ } PartOf(?x, ?c) \rightarrow (.5)\ Male(?x)$$
$$Couple(?c) \text{^} PartOf(?m, ?c) \text{^} Male(?m) \text{ ^ } PartOf(?f, ?c) \rightarrow Female(?f).$$

(One slot filler of a couple is male and the other is female.)

As we have noted, one of the problems with such structures has been that they have exceptions. This can be straightforwardly dealt with by using probabilities near, but less than, 1 on the constraints characterizing a structure. This high probability biases inference according to the information in the structure while permitting exceptions.

A key aspect in using scripts and frames is the matching process. For example, in a typical ap-

proach to resolving the pronominal references in "the couple went for a walk, he held her hand", the referent of "he" and the possessor indicated by "her" are matched to the male and female members of the couple frame. In our approach, these matches are represented by identity propositions. To say that the referent of "he" is the male of the couple is to say that they are identical, i.e., $Same(heRef, maleSlotFiller)$.

In this approach, therefore, the procedural problem of matching an object to a slot becomes a factual question about identity. As will be illustrated in the next section, this helps explain how the full range of a person's knowledge and inference abilities can be used to find the best filler for a slot. This is a much harder task when matching is a procedural matter that is conducted by a separate algorithm or subsystem from inference about the world.

Finally, we need one more constraint to infer that matches are made at all. For example, the constraints mentioned thus far would not favor a world where a pronoun has an antecedent, e.g., where $heRef$ is identical to some other object introduced into the discourse. We can favor such worlds with the following "minimal interpretation" constraint: $True() \rightarrow .51\ Same(?x, ?y)$.

All else being equal, this reduces the probability of worlds where two objects are not equal. Of course, other constraints, e.g., that "he" typically refers to a male, can override this bias.

This is, of course, a gross oversimplification. A much richer set of constraints are involved in favoring interpretations with reference, but this will be sufficient for our purposes.

4.3 Language understanding as a MAP inference problem

It is now possible to somewhat more precisely characterize the language understanding problem within the inferential approach. The listener's knowledge is characterized by the set of constraints, C = U ∪ SYN ∪ SEM ∪ PRAG ∪ WORLD, i.e., the union of knowledge of the specific utterances made together with linguistic and nonlinguistic knowledge. The goal of listening is characterized as finding the most likely world given this knowledge. This world includes the identity relationships characterizing the references of phrases in U. For example, if one of the utterances encoded in U is "John likes Mary because she is funny", the most likely world will have the statement $Same(sheRef, mary)$.

More technically, this characterization treats language as a maximum a posteriori inference (MAP) problem. This does not however fully capture the listener's situation. For example, the case where the most likely interpretation of a sentence has 99% probability is different from the situation where the most likely interpretation has 33% and the next most likely has 31%. In the work presented here, it will be sufficient to illustrate the benefits of the inferential approach by treating understanding as a MAP problem, although future work will need to address this issue.

5 Analyses

We now demonstrate how the substrate approach enables a unified analysis of the difficult kinds of utterances that motivated this investigation.

This is a new incarnation of the inferential approach and thus many aspects of it are oversimplified and provisional. In particular, many of the analyses below rely on simplified constraints that use probabilities that are at present guessed at. These were adequate and necessary for the goal of this work, namely to begin to develop an approach that provides a unified treatment of many difficult aspects of language. Once the outline of an approach exists, it will then become possible to more carefully elaborate aspects of the theory.

5.1 Nonlinguistic inference

We begin first by illustrating how cases where nonlinguistic inference help disambiguate a reference. Consider:

(11) John paid Fred for the car he gave him.
(12) John paid Fred for the car he wrecked.

In the most likely interpretation of (11), Fred ("he") sold the car to John. In (12), John ("he") wrecked Fred's car and compensated him. Constraints such as the following capture the relevant knowledge:

$$Give\ (?x, ?y) \rightarrow 1\ Pay(?y, ?x)$$
$$Damage(X, Y)\ +\ Owned(X, O) \rightarrow .8\ Pay(X, O)$$

In each sentence, there are two possible referents for "he" in the text: John ($Same(heRef,$

$john$)) and Fred ($Same(heRef, Fred)$). In the first sentence, there is a world where Fred is the referent. In this case, the commercial transaction would explain the paying of the money. In the world John is the referent, then John giving Fred a car would not explain John paying Fred, the paying event would be unexplained and that world would have a lower probability. Thus, we infer that Fred is the referent. A similar pattern of reasoning yields the correct referent in the second sentence.

In general, each possible identity relation will imply a possible world. Since world and linguistic knowledge are represented using constraints, they both jointly determine a probability for that world. The best referent is the one that is true in the world with the highest probability.

5.2 Implicit co-referent

Although we deal with several kinds of implicit reference (e.g., references from common knowledge, members of sets and sets composed of past elements in discourse), it is possible to give them a unified treatment. In each of these cases, world knowledge licenses the inference of *implied objects* not explicitly referred to in the utterance. Then the minimal interpretation constraint favors possible identities between the explicit referent and the implied objects. Finally, world knowledge helps rank these identities according to their likelihood. The following cases illustrate this chain of inference.

Referent is member of set. Consider (3). As described in section 4, "couple" licenses the inference of two entities who are likely to be a male (m) and a female (f). "he" liscences a male (hm) and "she" a female (sf).

There are several worlds based on combinations of identity propositions. These are a few:

1. $m = fm;\ f = hm$. "he" refers to the female member of the couple and "she" to the male.
2. $m = hm;\ f = fm$. "he" and "she" refer to the male and female member of the couple respectively.
3. $m = hm;\ f \neq fm$. "he" refers to the male of the couple but f refers to someone not mentioned in the couple.

... .

All worlds (e.g,. world 1) where "he" and "she" refer to people of the wrong gender are given very low probability because of the conditionals describing the semantics of the pronouns. Worlds where one of the pronouns do not refer to something explicit or implicit in the discourse (worlds 3 onward) have their probability decreased by the minimal interpretation constraint. World 2, where "he" and "she" refer to the male and female of the couple respectively is the only one that does not violate the semantic and minimal interpretation constraints and thus is the one with the highest probability.

A more complicated variation of (3) is (13):

(13) The couple went for a walk with their daughter. They held her hand.

There are two antecedent females for "her", the daughter and the female member of the couple, although the case where "her" refers to the "daughter" is clearly more likely. Such examples pose severe difficulties for algorithms used in structured knowledge approaches. Properly matching "her" to the female slot of the couple frame requires ruling out the match with daughter based on a chain of inference involving the fact that people tend not to hold their own hands and matching the referent of "her" to the female member implies that she is holding her own hand. Matching algorithms generally do not themselves make such inferences and are difficult to integrate with algorithms that do.

In the inference approach, one needs simply add a constraint indicating that people tend to hold other people's hands:

$$HoldHand(?x, ?y) \rightarrow .98\ \neg Same(?x, ?y).$$

Adding this constraint makes the interpretation where "her" refers to the female member of the couple less likely and thus the most likely world will have "she" refer to the daughter.

Composed referent. In (4), "both" refers to the set composed of the home and visiting goalies. The analysis here is similar to the previous case. "Both" implies the existence of a set with two objects, $o1$ and $o2$. The minimal interpretation constraint favors worlds where $o1$ and $o2$ are identical to elements in the discourse. This leaves possible worlds with two different sets of assumptions:

$o1$ = visiting goalie; $o2$ = home goalie or
$o1$ = home goalie; $o2$ = visiting goalie.

Each possibility is equally likely and thus there will be a tie for the most likely world. Ideally, there would be some way of inferring the equivalence of these worlds and thus in some sense collapsing them into one world or interpretation. However, even under the present circumstance, in each of the most likely worlds, the goalie NPs refer to members of the set "both" refers to.

Part of event. In (5), "the thunder" refers to the thunder that was a part of the storm mentioned in the first sentence. If the listener knows the following constraint, i.e., that thunder tends to be part of storms, then the finding the referent is simple.

$$Storm(?\,s) \rightarrow (.7)Thunder(?\,t)\,\wedge \\ PartOf(?\,t, ?\,s), ?\,t$$

"The storm" licenses the inference of a storm, s, and the constraint licenses the possible existence of thunder, t, that is part of the storm. "the thunder" licenses the inference of thunder ($thunder$). The minimal interpretation constraint favors worlds where "$Same(t, thunder)$", i.e., the interpretation that "the thunder" refers to the thunder that is part of "the storm". As in the previous cases, the interpretation follows directly from the meaning of phrases and the minimal interpretation constraint.

Verbal antecedent. The analysis of (6) is nearly identical. The only difference is that the thunder is inferred from the mention of the thundering event.

Common Knowledge. In (7), George Bush is not introduced or implied previously in the utterance. However, for most people his existence is known and hence part of the C via $WORLD$. Thus, "George Bush" licenses the existence of a person whose name is "George Bush" and the minimal interpretation constraint favors worlds in which the referent of "George Bush" is identical to the George Bush of common knowledge.

5.3 Non-literal reference

Cases of non-literal reference can also be dealt with using a combination of identity matching and the minimal interpretation constraint. In this case, non-literal referents are identified on the basis of their relation to the literal referent.

To illustrate, in (9), while the ham sandwich did not order the coffee, the person who did was related to that ham sandwich (by virtue of having ordered it). This suggests that in cases where the actual referent is not the literal referent, that there is often nevertheless a relation between them. If so, then many non-literal references can be understood by first determining the literal reference to be unlikely and then searching for an object related to it that can plausibly be the actual referent.

We represent the possibility of non-literal references being related to literal references with the "related referent constraint":

$$LitRef(?\,w, ?\,litRef) \rightarrow p_{nonlit} \\ Ref(?\,w, ?\,ref)\,\wedge\, ?\,R(?\,ref, ?\,litRef).$$

This constraint is similar in spirit to formula occurring in Pustejvsky's (1995) treatment of coercion, though in its present manifestation, it is used to explain a wider range of other phenomena.

Treatments of coercion and other phenomena try to limit the set of relations that can be involved in these phenomena. Although we take no stance on the content of such sets, constraints can be formulated to restrict the set of relations that can relate literal and non-literal referents.

We can deal with several cases of non-literal reference. Each analysis only involves the few very general constraints about reference already discussed, the literal semantics of each utterance and some world knowledge. No special provision need be made for each phenomenon. Much variation is accounted for by non-linguistic "context".

Coersion. In coercion, a phrase appears in a position that calls for a different type of referent then the actual referent of the phrase. For example, in (8) and (14), "begin" requires an event or action, while book is an object.

(14) The student began the book.

In (8) the action is the writing of the book while in (14) it is the reading of the book. In both cases, the literal referent is "coerced" from being of type object to type action. These also illustrate that coercion can be ambiguous and context-sensitive.

Coercion and the disambiguation of coercion are straightforwardly explained by the related referent constraint. We call the actual referent of "the book", $bookRef$, and the literal reference, $bookLitRef$. The semantics of "began" constrains the category of book-ref to be an action ($IsA(bookRef, Action)$). Thus, worlds where the actual reference of "the book" is the literal refer-

ence (i.e., $Same(litRef, bookLitRef)$) have very low probability because the category of literal reference is an object ($isA(bookLitRef, object)$) and not an event. (This presumes background knowledge that books are not objects).

Only worlds where the actual referent is a non-literal referent (that is, related (by relation R) to the literal referent) remain. At least two actions are related to books: reading and writing.

$Author(?p) \rightarrow (.2) Write(?p, ?b) \wedge$
$Book(?b), ?b, ?w$
(Some authors write books.)
$Student(?p) \rightarrow (.9)\ Read(?p, ?b) \wedge Book(?b)$
(Most students read books.)
$Students(?p) \rightarrow (.99) \neg Author(?p).$
(Most students are not authors.)

Since students are more likely to read books than write them, the world where the student began reading ($Same(R, Read)$) the book is favored. Likewise, since authors write books, the world were the actual referent is a writing event ($Same(R, Write)$) is more likely in (8).

Metonymy. The account of metonymic references is almost identical. Worlds where the literal referent is the actual referent have low probability because they clash with world knowledge. Thus, a metonymic reference is more probable.

For example, with respect to (9), ham sandwiches do not order coffee. Thus, "ham sandwich" cannot refer to the ham sandwich and must instead refer to something related to it. There are many things related to the ham sandwich: e.g., the chef, the plate, the ham in it, and the waiter who served it. However, since these order coffee infrequently, if ever, worlds where they are the actual referent have low probability. Since customers do often order coffee, the world where the customer is the referent is most likely.

6 Conclusions

Language use involves difficult problems, many of which are manifest in resolving ambiguous, implied and non-literal references. These problems seem to involve the interaction of linguistic and nonlinguistic factors. Our approach attempts to deal with these problems by framing language understanding as an action understanding inference problem. This enables a unified treatment of phenomena that to our knowledge have not yet been given a single explanation. It accounts for subtle variations in reference judgments based on nonlinguistic context. This work differs from past inferential approaches by extensively using identity constraints and thereby enabling an account of non-literal references, which had not been heretofore possible in inferential frameworks.

Fully realizing this approach will require a much more linguistic and non-linguistic knowledge. Acquiring it will involve learning from many instances of actual utterances. The preceding analysis provides a target for this effort and suggests significant benefits would result from it.

References

Clark, H. (1996). Using Language. New York, NY: Cambridge University Press.

Hobbs, J. R., Stickel, M. E., Appelt, D., & Martin, P. (1990). Interpretation as Abduction (No. 499). Menlo Park, California: AI Center, SRI Internationalo. Document Number)

McShane, M. (in preparation). The Requirements of Robust Reference Resolvers.

Milch, B., Marthi, B., Sontag, D., Russell, S., Ong, D. L., & Kolobov, A. (2005). BLOG: Probabilistic Models with Unknown Objects. Paper presented at the IJCAI-05, Edinburgh, Scotland.

Minsky, M. L. (1975). A Framework for Representing Knowledge. In P. H. Winston (Ed.), The Psychology of Computer Vision. New York, NY: McGraw -Hill.

Mitkov, R. (2000). A new fully automatic version of Mitkov's knowledge-poor pronoun resolution method. Paper presented at the CICLING-2000.

Murugesan, A., & Cassimatis, N. L. (2006). A Model of Syntactic Parsing Based on Domain-General Cognitive Mechanisms. Paper presented at the 8th Annual Conference of the Cognitive Science Society, Vancouver, Canada.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: polysemy. Linguistics and Philosophy, 3(1).

Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, MA: MIT Press.

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum.

A Word-Probabilistic Interface to Dialogue Modules

Alex Chengyu Fang, Weigang Li and Jonathan Webster
Department of Chinese, Translation and Linguistics
City University of Hong Kong
83 Tat Chee Avenue Kowloon, Hong Kong
{acfang,weiganli,ctjjw}@cityu.edu.hk

Abstract

A telephony dialogue system is described that performs speech-driven terminological translation. In particular, a novel approach is presented and discussed that is designed to probabilistically choose from a set of predefined, plan-based dialogue modules in order to maximise system usability. It is shown that words of different lengths, defined in terms of characters and syllables, demonstrate predictable degrees of recognition accuracy by the ASR engine. When expressed probabilistically, such varying degrees can be effectively used for the choice of appropriate dialogue modules. The novelty of this work is the measurement of word correct rate (WCR) as a function of grammar size and word length, expressed as WCR based on characters (*WCR-C*) and WCR based on syllables (*WCR-S*). The experimental results show that *WCR-C* and *WCR-S* can offer strong support in the development of an effective dialogue system, enhance dialogue flow and improve usability.

1 Introduction

Man-machine dialogue systems make use of different dialogue strategies to clarify user intent and to respond in an appropriate way. Typically, a dialogue system comprises different dialogue modules that handle different situations in the process of intension clarification. In speech-driven systems in practice, this boils down to the accuracy of the automatic speech recognition (ASR) system and how the system responds to different situations. For example, given the following dialogue turn:

System: Which term would you like to translate?
User: Gearbox.

the ASR engine will have a Boolean return. In the case of a positive one, the dialogue system will respond:

System: You said 'gearbox'. Its translation in Chinese is 齿轮箱.

With a negative ASR return, the system will say something like:

System: I'm sorry. Could you please repeat?
User: I said gearbox.

To enhance system usability, a third scenario is often necessary, where the caller is asked to confirm the ASR return:

System: Did you say gearbox?
User: Yes.

As can be seen, the three dialogue modules are components of an interactive session that attempts to verify the semantics of caller intent. A spoken dialogue system is typically configured to make use of the confidence level provided by the ASR engine in order to decide which module to opt for. There is also work to combine a second confidence score that represents an estimation of the mapping between the ASR result and user intention.

In this article, we report our work that aims to establish a third confidence score that is estimated externally on the linguistic string uttered by the speaker. Simply put, the score is an estimation of the probable ASR error rate according to the length of the word. The proposal of this additional confidence score is necessary since the other two scores do not take into account the

fact that words of different lengths tent to have a different impact on the ASR engine. In addition, the size of ASR language models or grammars also has a significant impact on ASR performance. Our work to be reported here is therefore concerned with ASR evaluation according to two parameters: word length and grammar size.

Effective evaluation is an important task in spoken language dialogue systems (SLDS). Generally speaking, there are two purposes. One is to compare performance of different systems. The other is to improve the evaluated system itself. Different methodologies have been proposed to evaluate components in spoken language dialogue systems, such as Word Error Rate (WER) and weighted keyword error rate (WKER) (Nanjo and Kawahara, 2005; Hildebrandt et al., 1996). Higashinaka and colleagues describe a method for creating an evaluation measure for discourse understanding in spoken dialogue systems (Higashinaka et al., 2004). There is also a focus on user-related issues, such as user reactions to SLDS, user linguistic behaviour or major factors which determine overall user satisfaction (Walker et al., 1997; Walker et al., 2001; Hartikainen et al., 2004). There is increased focus on usability evaluation of SLDS in recent years (Dybkjr and Bernsen, 2001; Park et al., 2007) and metrics have been proposed, such as modality appropriateness, naturalness of user speech, and output voice quality.

All these methods are concerned with objective or subjective criteria of SLDS (Larsen, 2003). They aim to describe the system performance on the whole or part. Additionally, they all evaluate SLDS beyond the word level. This article discusses the fine evaluation of word-level performance in terms of word correct rate (WCR) and argues that there is much useful information at the word level that can improve SLDS performance effectively and efficiently.

2 Motivation

RAMCORP is a project that aims at the design and construction of a telephony dialogue system that provides on-the-spot machine translation of terminologies of a pre-defined domain. The interactive dialogue system uses Nuance, an off-the-shelf automatic speech recognition system, for the recognition of key words. In order to maximize transaction completion rate, RAMCORP will consist of several dialogue modules with different dialogue turns. A novelty of the project is to dynamically determine which dialogue to opt for according to the word being recognized. To achieve this, empirical experiments were carried out to ascertain the word correct rate (WCR) according to grammar size and word length. While it is common practice to measure WCR according to grammar size, the measurement of WCR as a function of word length has not been widely reported before. We define word length in two different ways: according to number of characters (WCR-C) and according to number of syllables (WCR-S). Results of the empirical experiments will ultimately inform the design of a formula that dynamically calculate the likelihood of a word being correctly recognized according to the three parameters, i.e., grammar size, number of characters, and number of syllables. Effectively, the system will be able to predict the likelihood of a word being correctly recognized and choose a corresponding dialogue module according to this likelihood.

This paper will focus on the empirical experiments that were carried out to establish the baseline statistics for Nuance. It will first of all report data selection including the selection of participating subjects and the selection of words that were used to form mock-up grammars of various sizes. It will then evaluate the ASR performance and report the resulting WCRs according in and discuss major findings.

3 Experiments and Analysis

3.1 Experimental setting

The off-the-shelf application used in this paper is Nuance Voice Platform (NVP). A demo dialogue system with word grammar rules is built for evaluation. Four grammars were constructed, consisting of only words to be recognized without any context cues. They respectively include 500, 1000, 2000 and 4000 words randomly selected the machine readable Collins English Dictionary. Twenty subjects as evaluators were

invited to participate in the experiment. Each was asked to read four groups of 50 words randomly selected from the four grammars.

We thus obtained 20 sets of recognition results for grammars of four different sizes. The results of the experiment are summaries in Table 1.

S	WCR_{500}	WCR_{1000}	WCR_{2000}	WCR_{4000}	M
1	68.0	60.0	60.0	48.0	59.0
2	48.0	62.0	44.0	44.0	49.5
3	64.0	70.0	62.0	52.0	62.0
4	78.0	84.0	64.0	62.0	72.0
5	72.0	64.0	66.0	60.0	65 .5
6	62.0	60.0	46.0	44.0	53.0
7	84.0	58.0	58.0	50.0	62.5
8	88.0	66.0	76.0	64.0	73.5
9	72.0	80.0	56.0	50.0	64.5
10	68.0	52.0	58.0	58.0	59.0
11	64.0	64.0	56.0	50.0	58.5
12	74.0	60.0	50.0	40.0	56.0
13	58.0	58.0	64.0	54.0	58.5
14	72.0	44.0	66.0	44.0	56.5
15	82.0	74.0	76.0	50.0	70.5
16	82.0	78.0	82.0	58.0	75.0
17	76.0	72.0	74.0	56.0	69.5
18	82.0	84.0	62.0	58.0	71.5
19	78.0	68.0	68.0	68.0	70.5
20	76.0	70.0	76.0	58.0	70.0
M	72.4	66.4	63 .2	53.4	63.85

Table 1: Word correct accuracy and grammar size

3.2 Evaluation of WCR on Grammar Size

The most popular evaluation metric of ASR is Word Error Rate (WER), which is the minimum string edit distance between the correct transcription and the recognition hypothesis. There will be some new measures to propose to finely evaluate the dialogue system. In order to distinguish traditional WER, Word Correct Rate (WCR) is defined in this paper:

$$WCR = \frac{Count(Correct)}{Count(Total)} \quad (1)$$

Count(Correct) is the number of words recognized correctly, and *Count(Total)* is the total number of words to be recognized. WCR describes the performance of dialogue system with a certain number of grammar rules. The average *WCR* of the system with four different grammar scales is called WCR_a. It can be calculated:

$$WCR_a = \frac{\sum_{scale \in SSet} Count(Correct)}{\sum_{scale \in SSet} Count(Total)} \quad (2)$$

$$SSet = \{500,\ 1000,\ 2000,\ 4000\}$$

The average WCR of twenty evaluators on the system with certain scale grammar rules is called WCR_{sca}, which can be calculated through the following formula:

$$WCR_{sca} = \frac{\sum_{i=1}^{n} WCR(i)}{n} \quad (3)$$

The number *n* is the number of evaluators. There are twenty persons to participate in our experiments.

The evaluation results show that dialogue system has different recognition performance with different grammar sizes. According to Figure 1, the observable trend is that there is a consistent reduction of system performance with increased grammar size.

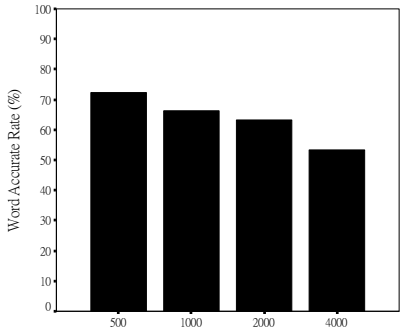


Figure 1: Word accurate rate and grammar size

Figure 1 shows that recognition accuracy drops from 72.4% to 53.4% with a mean of 63.85% when grammar size is increased from 500 to 4000. This observation suggests the need to improve system performance by using dynamically constructed hierarchical grammars instead of monotonic grammars for every recognition slot. Dynamically constructed hierarchical

grammars are different from monotonic grammars in that grammar rules are typically classified into several groups according to their prior probabilities to be recognized. The prior probabilities can be obtained from context and other related information. How to get operable hierarchical grammars will be an important part of our future work on RAMCORP.

See Figure 2 for system performance with the 20 subjects.

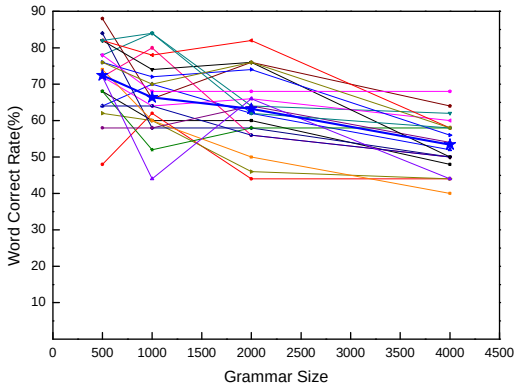


Figure 2: System performance with the 20 subjects in the experiment

There is considerable fluctuation in WCR for the 20 subjects with a standard deviation of 7.56, as demonstrated in Figure 2, which is expected for a telephony dialogue system. It should be noted that the twenty evaluators are non-native English speakers from China so the actual WCR of the evaluated system would be higher than the WCR values required in our experiments if the callers were native speakers requesting the translation of terminologies from English to Chinese.

Figure 3 shows that, across the four grammars on average, the system had varying degrees of performance with the 20 subjects. The maximum is 75.0% and the minimum 49.5% with a mean of 64.1. The standard deviation is 7.56%. Such variations are expected for a telephony dialogue system open to a wide range of speaker diversity.

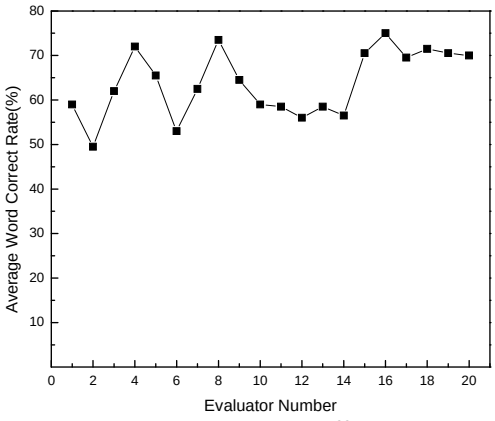


Figure 3: Average WCR with different evaluators

3.3 WCR Variation and Word Length in Characters

Word length defined in number of characters is the second parameter concerned in this study that is expected to have an impact on recognition performance. The WCR based on character length is called WCR_{cl}. It is an average value calculated according to Equation (4):

$$WCR_{cl} = \frac{\sum_{scale \in SSet \ \& \ i \in ESet} Count(Correct_{cl})}{\sum_{scale \in SSet \ \& \ i \in ESet} Count(Total_{cl})} \qquad (4)$$

where, SSet is scale set which represents the same meaning in Equation (3). ESet is the evaluator set {1, 2, 3, ..., 20}. $Count(Correct_{cl})$ is the correctly recognized number of words with length "character length (abbr. cl)". $Count(Total_{cl})$ is all test words which length is equal to cl. The evaluation results are summarized in Table 2. The second column in Table 2, marked *Test Set*, lists the word length distribution of all the test words randomly selected in the experiment with # indicating the actual number of words selected and % its proportion in all of the test words selected. The third column, *Lexicon*, is the distribution of all words in the dictionary with # indicating the total number of words of the concerned length and % the proportion of such words in the dictionary.

It can be seen that the word length varies from 1 to 21 characters and that the selected

words in the test set form a good representation of those in the lexicon in terms of distribution of character lengths. Words with lengths between 4 and 12 characters account for about 90 percent of total number.

C	Test set		Lexicon		WCR_{cl}
	#	%	#	%	
1	6	0.15	32	0.06	50.00
2	4	0.10	248	0.46	75.00
3	79	1.98	841	1.56	35.44
4	218	5.45	2399	4.45	53.67
5	320	8.00	3995	7.41	49.06
6	471	11.77	5958	11.05	58.81
7	588	14.70	7187	13.33	61.22
8	528	13.20	7554	14.01	61.36
9	572	14.30	7306	13.55	71.15
10	456	11.40	6066	11.25	72.15
11	313	7.83	4448	8.25	70.93
12	177	4.42	3133	5.81	71.19
13	136	3.40	2043	3.79	75 .74
14	59	1.47	1240	2.30	64.41
15	38	0.95	744	1.38	73.68
16	18	0.45	388	0.72	77.78
17	9	0.22	216	0.40	66.67
18	5	0.13	81	0.15	100.00
19	2	0.05	38	0.07	100.00
21	1	0.03	5	0.01	0.00
M	4000	100.00	53916	100.00	63.85

Table 2: WCR based on character length

As Figure 4 clearly shows, words with different character lengths have different impact on system performance as suggested by WCR_{cl}. It can be observed from the graph that there are some ups and downs at the two ends of WCR_{cl}-*length* curve. This phenomenon can be caused by two possible reasons. Firstly, words shorter than 4 and longer than 12 characters in length are relatively small in population. The randomly selected few cannot support statistic results sufficiently. Secondly, the evaluators involved in these experiments are non-native speakers of English while all the test words were selected randomly from a large dictionary. Therefore there were unfamiliar words for the evaluators, which resulted in inaccurate pronunciations and subsequently recognized inaccuracies by the system.

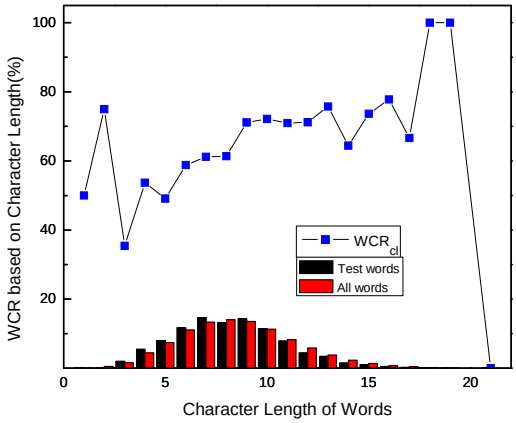


Figure 4: WCR based on character length

But the predominate words with lengths between 4 and 12 have a consistent trend and WCR_{cl} increases steadily with word length. Generally, the longer a word is, the more likely the word is accurately recognized. One observation is that words between 6 and 8 characters in length have a similar WCR while those between 9 and 12 have a similar but higher WCR. This suggests that the use of word character as a measurement unit has a wide range of variation in terms of WCR, which calls for the use of another measurement unit that exhibits a lower degree of variation. As a result, we introduced the use of syllables as a second measurement unit, to be discussed in 3.4 below.

Based on the evaluation results of WCR_{cl}, a more suitable dialogue model can be designed for improving performance of dialogue systems. Simple dialogue modules can be applied to recognize long words because these words have a relatively high WCR_{cl}. Conversely, complex dialogue modules with extended interactive turns will be needed for shorter words that typically have a lower WCR_{cl}. By doing so, a dialogue system with a good balance between conciseness and accuracy can be achieved.

3.4 WCR Variation and Word Length in Syllable

As mentioned above, words of different lengths have different impact on system performance

measured in WCR_{cl}. In fact, the major factor can be attributed to syllable information, which influences the accuracy of word speech recognition significantly. In this sense, the number of syllables of a word may demonstrate more precisely the correlation between word length and recognition accuracy.

For this purpose, a machine-readable pronunciation dictionary was used to retrieve the number of syllables for each of the test words selected for the experiment. The WCR based on syllable length, WCR_{sl}, is calculated by the following formula:

$$WCR_{sl} = \frac{\sum_{scale \in SSet \ \& \ i \in ESet} Count(Correct_{sl})}{\sum_{scale \in SSet \ \& \ i \in ESet} Count(Total_{sl})} \quad (5)$$

The formula is similar to Equation 4. The only difference between them is that $Count(Correct_{sl})$ is the word count with syllable length "sl" being recognized correctly. The WCR_{sl} results are listed in Table 3.

S	Test set		Lexicon		WCR_{sl}
	#	%	#	%	
1	429	10.73	4028	0.06	50.00
2	1283	32.07	15582	0.46	75.00
3	1103	27.58	15501	1.56	35.44
4	775	19.38	11020	4.45	53.67
5	293	7.32	5322	7.41	49.06
6	91	2.27	1871	11.05	58.81
7	19	0.47	507	13.33	61.22
8	7	0.18	86	14.01	61.36
M	4000	100.00	53916	13.55	71.15

Table 3: WCR based on syllable length

The first column S shows the word length in terms of syllables. The second column in Table 3 is the syllable length distribution of all test words with # indicating the actual number of words selected and % the proportion of such words in the total number of test words. The third column, marked Lexicon, is the distribution of all words in the machine-readable pronunciation dictionary. # indicates the actual number of words of a certain length and % the proportion of such words in the lexicon. As can be seen from the table, the selected words and the lexicon show good similarity in terms of distribution, suggesting that the test data are sufficiently representative. Words of up to 6 syllables in length make up more than 99 percent of the total test set with a small margin of proportion for words with 7 syllables or more.

Figure 5 is a graphical representation of Table 3. It can be observed that WCR_{sl} for words with less than 7 syllables shows a consistent rise as a function of syllable number, increasing steadily together with the increase of word length measured in terms of syllables.

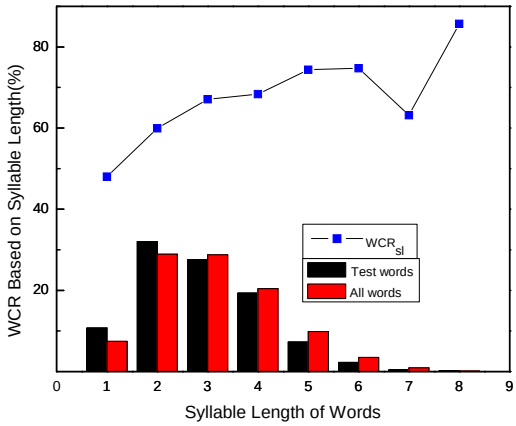


Figure 5: WCR based on syllable length

Compared the results with Figure 4, we can determine that the WCR_{cl} jump from 5 characters to characters is because words with 5 characters and 6 characters will have different syllables which influence the accuracy of their speech recognition. A similar phenomenon happens in 8 characters and 9 characters in Figure 4. The evaluation results offer support for designing an effective dialogue system.

4 Conclusions

This paper presented an experiment to evaluate the performance of Nuance for its recognition accuracy measured in word accurate rate (WCR). While conventional measurement is typically conducted in conjunction with grammar size, we designed a novel approach to measure WCR as a function of word length measured in terms of characters and syllables. Results show that while WCR drops with the

increase of grammar size, there is also the tendency for WCR to rise as a function of word length. Between characters and syllables, the experiment demonstrated that the latter is a better indication of the correlation between WCR and word length.

The results confirms the conventional wisdom in the first place that, instead of using a monotonic grammar which tends to be large in size and therefore affects WCR, a hierarchical grammar generated dynamically should be preferred for better WCR. This raises an interesting suggestion for the RAMCORP project to augment the list of terminologies in such a way that they can be effectively sub-classified in order to reduce recognition space and therefore to increase WCR. Secondly, the results suggest that better system performance can be expected when RAMCORP moves into a stage that involves the recognition of longer terminological phrases.

The most significant suggestion from the experiment is that a dynamically constructed dialogue model can be possibly achieved based on the word returned by the recognition slot. Such a model can be driven by a probabilistic engine that considers grammar size and word length measured in characters and syllables. Within such a probabilistic dialogue model, modules with different interactive turns can be selected according to the word recognized and returned by the system. While the general principle is that shorter terminologies require more dialogue turns to achieve a completed transaction, the system can be fine tuned for even better transaction completion rate based on probabilities associated to each keyword in the grammar. Such a dialogue system will require a self-maintenance mechanism of the grammar that updates itself for recognition probabilities for each individual rule.

On the basis of the suggestions above, future work will be carried out in two key areas: one is to construct effective hierarchical grammar rules using context and other features of the terminologies concerned in RAMCORP. The other is to design a probabilistic dialogue model for improving the usability of the service through maximally enhanced system performance. In addition, similar evaluation is required for the other languages involved in the project, including Chinese in the first instance and Korean and Japanese in the future.

Acknowledgments

This research was supported by the project "A Remote-Access Multilingual Corpus-Based System for Linguistic Applications (RAMCORP)", City University of Hong Kong. The authors would like to thank Mr John Poon and Ms Lanying Cheng of Nuance Communications, Hong Kong. The authors would also like to thank all the evaluators for their kind help with the experiments reported in this article.

References

Dybkjr, L. and N.O. Bernsen. 2001. Usability evaluation in spoken language dialogue systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems*, volume 9. pp 1–10.

Hartikainen, M., E. Salonen, and M. Turunen. 2004. Subjective evaluation of spoken dialogue systems using servqual method. In *Proceedings of ICSLP*. pp 2273–2276.

Higashinaka, R., N. Miyazaki, M. Nakano, and K. Aikawa. 2004. Evaluating discourse understanding in spoken dialogue systems. *ACM Transactions on Speech and Language Processing*, 1:120.

Hildebrandt, B., H. Rautenstrauch, and G. Sagerer. 1996. Evaluation of spoken language understanding and dialogue systems. In *Proc. ICSLP*, volume 2. pp 685– 688.

Larsen, L. B. 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In *Automatic Speech Recognition and Understanding.*

Nanjo, H. and T. Kawahara. 2005. A new ASR evaluation measure and minimum bayes-risk decoding for open-domain speech understanding. In *IEEE ICASSP*. pp 1053–1056.

Park, W., S.H. Han, Y.S. Park, J. Park, and H. Yang. 2007. A framework for evaluating the usability of spoken language dialog systems (sldss). *Usability and Internationalization.* pp 398–404.

Walker, M.A. C.A. Kamm, and D.J. Litman. 2001. Towards developing general models of

usability with paradise. *Natural Language Engineering*, (6). pp 363–377.

Walker, M.A., D.J. Litman, C.A. Kamm, and A. Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the 35th Annual Meeting of the ACL*. pp 271–280.

A Grammar Formalism for Specifying ISU-based Dialogue Systems

Peter Ljunglöf, Department of Linguistics, University of Gothenburg

We describe how to give a full specification of an ISU-based dialogue system as a grammar. For this we use Grammatical Framework (GF), which separates grammars into abstract and concrete syntax. All components necessary for a complete GoDiS dialogue system are specified in the abstract syntax, while the linguistic details are defined in the concrete syntax. Since GF is a multilingual grammar formalism, it is straightforward to extend the dialogue system to several languages.

The information-state update approach

The GoDiS dialogue manager [1] is based on formal semantic and pragmatic theories of dialogue, and provides general and fairly sophisticated accounts of several common dialogue phenomena such as interactive grounding (a.k.a. verification), accommodation, keeping track of multiple conversational threads, and mixed initiative. General solutions to general problems allow modularity, re-use and rapid prototyping.

GoDiS is based on the Information State Update (ISU) approach to dialogue management [4]. The ISU approach, which has been developed over the last 10 years in several EU-funded projects, provides a generalization over previous theories of dialogue management and allows exploring a middle ground between sophisticated but brittle research systems, and robust but simplistic commercial systems. In the ISU approach, a dialogue manager is formalized as: (i) an information state (IS) type declaration, (ii) a set of dialogue moves, and (iii) information state update rules.

In GoDiS, which is based on a theory of Issue-Based Dialogue Management (IBDM), a single script (called a *dialogue plan*) can be used flexibly by the dialogue manager to allow for a wide range of dialogues. The main benefit of the IBDM account as implemented in GoDiS is the combination of advanced dialogue management and rapid prototyping enabled by cleanly separating generic dialogue principles from application-specific domain knowledge.

GoDiS enables rapid prototyping of systems with advanced dialogue behavior. However, the GoDiS dialogue manager only communicates with the outside world using semantic representations called *dialogue moves*. The designer of the dialogue system must implement a translation between natural language utterances and dialogue moves, be it through a simple lookup table, or an advanced feature-based grammar. Furthermore, a speech-based system also needs a statistical language model or a speech recognition grammar. In this paper we show how a GoDiS dialogue system can be specified as a single grammar in the Grammatical Framework. All components necessary for a ISU-based dialogue system are then automatically generated from the grammar.

The GoDiS dialogue manager

The GoDiS system communicates with the user via *dialogue moves*. There are three main dialogue moves – requesting actions, asking questions and giving answers. Apart from the three main moves there are also different kinds of feedback moves – confirmations, failure reports and interactive communications management.

The basic building blocks in GoDiS are individuals, sorts, one-place predicates and actions. From these all necessay dialogue moves can be built, such as questions, answers, requests and feedback. To specify a GoDiS dialogue system, we have to give the following information: (i) the sortal hierarchy, (ii) the individuals and the sorts they belong to, (iii) the predicates and their domains, (iv) the actions, and (v) the dialogue plans. Furthermore, there has to be an interface to each external device.

Dialogue plans convey what the system can do and/or give information about. A dialogue plan is a receipt for the system, so it knows how to answer a specific question, or how to perform a given action. The dialogue plans can roughly be divided into three different kinds – actions, issues and menus. An *action plan* is when the user wants the system to perform an action, e.g., to call a contact in the address book. An *issue plan* is when the user wants the system to give information, such as telling the phone number of a contact in the address book. A special kind of action plan is the *menu*, where the user can select from any of a given number of sub-plans which the system then performs.

Grammatical Framework

Grammatical Framework [2] is a grammar formalism based on type theory. The main feature is the separation of *abstract* and *concrete* syntax, which is crucial for our treatment of dialogue systems. The abstract syntax of a GF grammar consists of declarations of categories and functions. Function declarations correspond to rules in a context-free grammar.

The concrete syntax consists of *linearizations* of the abstract functions. Linearizations are written in a typed functional programming language, which is very expressive but still decidable.

It is possible to define different concrete syntaxes for one particular abstract syntax, making GF a multilingual grammar formalism. Furthermore, the abstract syntax of one grammar can be used as a concrete syntax of another grammar, which makes it possible to implement grammar resources to be used in several different application domains. These points are currently exploited in the GF Resource Grammar Library [3], which is a multilingual GF grammar with a common abstract syntax for 13 languages, including Arabic, Finnish and Russian.

GoDiS specification as GF abstract syntax

Action and issue plans are specified as functions with result categories Action(m) and Issue(m) respectively, where m specifies which menu they belong to:

fun callContact : Name → Action(MakeCall)

fun searchForNumber : Name → Number → Issue(ManageContacts)

The first specification states that callContact is an action plan in the MakeCall menu. It takes one argument, which is the Name of the contact to call. The second specification is the issue plan searchForNumber. It also takes one Name argument, which the system will ask for if not already said by the user. The final Number argument represents the system's answer, and will be filled by the system when it knows the answer.

Everything else in the GF grammar specifies the ontology of the dialogue system. From the grammar we can extract the sorts and the sortal hierarchy, the individuals and the predicates.

Dialogue utterances as GF concrete syntax

The concrete syntax is responsible for translating everthing the user says into dialogue moves, and what the system might want to say into natural language. For each function in the abstract syntax, there has to be a corresponding linearization. E.g., the callContact action might have the following linearization:

lin callContact(x) = "call" ++ *variants*{x ; "a contact"}

This linearization will be used in several places by the dialogue system. First, we can use it in parsing the user utterances "call anna" (or "call a contact"): The result is the GF term callContact(anna) (or callContact(?)), which will be automatically translated into GoDiS dialogue moves. Second, the system will use it when presenting the MakeCall menu: "do you want to call a contact or call a number?" . Third, the system will generate different kinds of feedback moves containing the GF term: "so, you want to call a contact" or "I'm sorry, I cannot call a contact at this moment".

A short example

Assume that the user says "I'd like to call a contact please". This is recognized by the system as the dialogue move request(callContact), which means that GoDiS loads the associated action plan. In this plan it sees that it needs a value for callContact:Name, so it utters the dialogue move ask(?x.callContact:Name(x)). There is an extra linearization for callContact for handling wh-questions (not shown above), translating the dialogue move into "Which name do you want to call?". GoDiS incorporates the user answer as answer(Name(Kim)), and then tells the phone to look up Kim's number in the phone book and call. A confirmation dialogue move, confirm(callContact), is at the same time presented to the user.

References

[1] Staffan Larsson. *Issue-based Dialogue Management*. PhD thesis, Department of Linguistics, Gothenburg University, 2002.

[2] Aarne Ranta. Grammatical Framework, a type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189, 2004.

[3] Aarne Ranta, Ali El-Dada, and Janna Khegai. *The GF Resource Grammar Library*, 2006. Can be downloaded from the GF homepage `http://www.cs.chalmers.se/~aarne/GF`

[4] David Traum and Staffan Larsson. The information state approach to dialogue management. In Smith and Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer Academic Publishers, 2003.

Spoken Language Understanding in dialogue systems, using a 2-layer Markov Logic Network: improving semantic accuracy

Ivan V. Meza-Ruiz, Sebastian Riedel, and Oliver Lemon*
School of Informatics
University of Edinburgh
I.V.Meza-Ruiz,S.R.Riedel@sms.ed.ac.uk, olemon@inf.ed.ac.uk

Abstract

We describe a two layer Markov Logic Network (MLN) model for the Spoken Language Understanding (SLU) task in dialogue systems. We augment the set of features used in Meza-Ruiz et al. (2008) with the help of off-the-shelf resources. We show that this setup increases the performance of the previous MLN models, which also outperform the state-of-the art "Hidden Vector State" (HVS) model of He and Young 2006. In particular the 2 layer approach produces more accurate sets of slot-values for user utterances (9% improvement).

1 Introduction

The Spoken Language Understanding (SLU) task in dialogue systems consists in producing semantic representations for user utterances. In this work, our approach is trained on slot-value representations which are a common choice in the development of dialogue systems. Table 1 shows an example of slot-values as a semantic representation.

USER:*what flights are there arriving in Chicago on continental airlines after 11pm*
$GOAL$ =FLIGHT $TOLOC.CITY\_NAME$ =Chicago $AIRLINE\_NAME$ =continental_airlines $ARRIVE\_TIME.TIME\_RELATIVE$ =after $ARRIVE\_TIME.TIME$ =11pm

Table 1: Slot-values as a semantic representation.

This work is partially funded by EPSRC grant number EP/E019501/1 and the EC FP7 project "CLASSiC" (ICT-216594). www.classic-project.org.

In particular, we are exploring robust statistical models of the SLU task using the Markov Logic Network (MLN) framework(Richardson and Domingos, 2007). An MLN is a collection of weighted First Order Logic (FOL) formulae that serves as a template to instantiate complex Markov Networks (MNs). MLNs are particularly interesting for language modelling because they are easy extensible with new features and allow the use of complex relations between nodes of the networks. Figure 1 shows a MN for a slot-value representation. In this case, the lighter nodes represent the hidden variables (i.e., the slots to produce) while the darker nodes represent the observable variables (i.e., the words of the utterance).

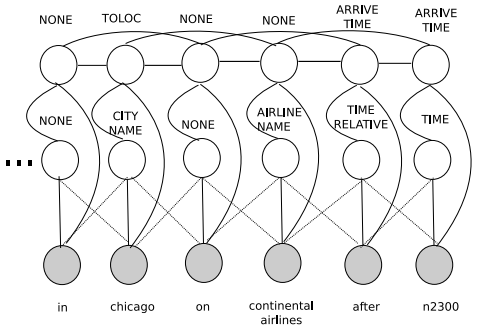


Figure 1: Markov Network as slot-values

In this paper, we focus on two aspects. First, the use of a two-layer MLN model to represent the slots. And second, the use off-the-shelf resources to extend the set of features available (e.g. POS taggers).

2 The MLN model

We split the SLU task into two. The first task consists into modelling the $GOAL$ element of the slot-

values. Figure 2 shows a MN for the goal of our example. You can notice, that the $GOAL$ element, depends on the whole utterance. The second task consists of modelling the rest of the slots. Figure 1 shows a two-layer MN for the slots of our example. The first layer can be seen as a named entity, while the second layer represents a modifier/function for those named entities.

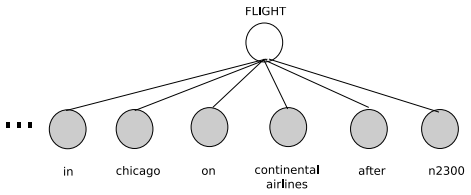


Figure 2: Markov Network for the goal slot

With the two layer model we aim to capture the relations between elements which constitute the slots. This is achieved by specifying the edges between the two layers using the FOL of the MLN. This model also includes the 1^{st} and 2^{nd} order Markov assumptions for the second layer. With these we aim to capture any dependency in the sequence of slots.

2.1 Feature extensions

In MLNs it is possible to add more observable variables which will be related to the input words. For this purpose, we use off-the-shelf resources to generate (i) POS tags for the words of the utterances using the TnT tagger (Brants, 2000), and (ii) syntactic chunks for the words of the utterances using the CASS chunker (Abney, 1996). With this information we define the following features for the slot model: Orthography and POS task of the word for a window of two previous and next words; a binary feature if the word is a number or the word is unknown ; and the head words of syntactic chunks.

3 Experiments and Results

For our experiments we use the Air Travel Information System corpus extended by (He and Young, 2006). This version is composed of slot-value labellings of the ATIS-2 and ATIS-3 training sets, and the ATIS-3 $NOV93$ testing set. We measured the *global score* and *exact match* metrics. This first metric measures the amount of slot-values recovered in the whole experiment, while the second one measures the exact set of slot-values recovered for each utterance.

The experiments tested the two layer models described in the previous sections. Table 2 presents the final results. Our baseline corresponds to the previous best result for the task, which outperforms the HVS model of (He and Young, 2006) on the labelling task (Meza and et al., 2008). In this case, the baseline is our starting point. The $MLN_{2-layer}$ model uses the features described in the previous section, plus a two layer model. The difference bewteen both models is statistically significant with $\rho < 0.05$.

$MLN_{baseline}$	Global	91.56%
	Exact	69.89%
$MLN_{2-layer}$	Global	92.99%
	Exact	78.97%

Table 2: f-scores for baseline and two layer model

4 Summary and discussion

We have shown an improvement on the performance of the SLU task by using a two layer model and by augmenting extra features in our previous model. In particular, the improvement of 9.08% in the *exact match* metric is interesting. This is because it shows that not only was the model able to identify more slot-values, but the set of slot-values for each utterance were more accurate.

References

Steven Abney. 1996. Partial Parsing via Finite-State Cascades. In *Natural LanguageEngineering*, volume 2(4), 337-344.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *ANLP*.

Yulan He and Steve Young. 2006. Spoken Language Understanding using the Hidden Vector State model, *Speech Communication*, volume 48(3-4), 262-275.

Ivan Meza, Sebastian Riedel and Oliver Lemon. 2008. Accurate Statistical Spoken Language Understanding from limited development resources, In *ICCASP08*.

Matt Richardson and Pedro Domingos. 2007. Markov Logic Networks. *Machine Learning*, volume 62, 107-136.

Negotiating spatial relationships in dialogue: The role of the addressee

Thora Tenbrink
Universität Bremen
SFB/TR 8 Spatial Cognition, Germany
tenbrink@uni-bremen.de

Elena Andonova
Universität Bremen
SFB/TR 8 Spatial Cognition, Germany
andonova@uni-bremen.de

Kenny Coventry
Cognition & Communication Research Centre
Northumbria University, UK
kenny.coventry@unn.ac.uk

Abstract

How do addressees who are not informed about targets contribute in a conversation to the negotiation of spatial locations? Results in dialogue research show the general importance of the addressee's reactions to a speaker's utterances. Results in spatial language research demonstrate the range of variability available to a speaker when providing a spatial description. In this paper, we combine these two approaches in order to investigate how the spatial position of objects in a dolls' house is negotiated, using a naturalistic dialogue scenario. Results show the ways in which the instructed person actively supports the negotiation of spatial reference, for example by pointing out ambiguities and suggesting alternative conceptual perspectives on the scene.

1 Introduction

When engaged in joint action, you may be asked to place an object in a particular position. How do you react? In theory, all you need to do is place the object and wait for the next instruction. However, in natural dialogue addressees do much more than that (e.g., Clark, 1996): they acknowledge the speaker's request, ask for clarification, or contribute to the description by expanding it or suggesting a different description. When placing objects, the situation becomes particularly complicated, as spatial terms can typically be interpreted in more than one way (e.g., Schober, 1993). Then how do addressees contribute to the given task of a spatial placement so that an agreement can be reached "well enough for current purposes" (Clark, 1996)?

Previous work on spatial language in dialogue has focused, for example, on direction-giving through a maze (Garrod & Anderson, 1987) or map task (Filipi & Wales, 2004), on spatial object reference, i.e., the identification of an object in contrast to other objects present in a scene accessible to both dialogue partners (e.g., Schober, in press), on route descriptions (Muller & Prévot, in press), and on the description of spatial relationships in pictures (Watson et al., 2004). In contrast, in a situation like the one just sketched, spatial language is used to instruct someone to place an object in a particular position. Such a situation involves a fairly strong knowledge discrepancy, raising the question whether the instructed person will be able to contribute any suggestions of their own at all. However, even at a brief glance at our data corpus (targeting such a scenario), we encounter the following exchange:

director: also oben links in dem Ba+ in dem Zimmer also oben links [okay, at the top and left in the ba+ in the room that is at the top left]
matcher: neben dem Fenster die Dusche?
[next to the window the shower?]
director: ne - mehr rechts also im Raum hinten rechts
[no - more right that is in the room at the back right]

Apparently, the matcher has quite a good idea already of where the shower (a piece of dolls' house furniture) is to be placed, and makes an informed suggestion, which is taken up and corrected by the director. Strikingly, the director's subsequent description departs fundamentally from the original: obviously, the conceptual perspective on the scene

has changed by the matcher's utterance. This phenomenon can be regarded as a specific case of the generally well-documented dialogue processes of repair, clarification, and grounding (e.g., Clark & Krych, 2004); or in Schegloff's (1997) terms, 'candidate understandings' or 'appendor questions'. In this paper we ask how the peculiarities of spatial language come into play in this kind of collaborative negotiation procedure.

Spatial language constitutes a common class of natural language that is particularly regularly used in everyday discourse (Talmy, 2000). This includes the so-called projective terms which indicate a direction (*left, right, above, below, front, back*); these are interpreted against conceptual reference systems (Levinson, 2003; Tenbrink, 2007) and may be relevant in a static or a dynamic sense (van der Zee and Slack, 2003). Further, there are topological terms which indicate aspects of contiguity (*on, in, at*), path-related terms (e.g., *across, through, along*), distance-related terms (e.g., *near, far, close*), and others. These terms can be used in a broad variety of discourse tasks and then exhibit different features and implications (Bateman et al., 2007). For instance, searching a hidden object in a small-scale array requires descriptions on a different level of granularity than describing a route for a stranger in town. Furthermore, spatial terms may be used in order to describe an object for reference purposes (e.g., *the car with the blue top*), or in order to describe an object's location (e.g., *the car is in front of the house*). To describe an object's position in yet more detail, one may wish to describe the orientation of an object, again using spatial terms (e.g., *the front of the car points towards the house*). Here we focus on a scenario designed to maximise the occurrence of spatial terms by combining the latter three options in an *object placement* task. To reach that goal successfully, speakers must agree on an object's identity as well as its location and orientation.

Much work on dialogue (e.g., Garrod & Anderson, 1987) relates to spatial settings, but the intricate repertory of how to describe spatial relationships has not yet been explored thoroughly with respect to dialogue phenomena. Crucially for our present interests, little is known yet about how the *addressee* contributes to establishing spatial relationships in task-oriented dialogue, since the focus of attention in most analyses mostly lies either on the direction giver or on particular discourse processes such as interactive alignment (Pickering & Garrod, 2004). Our present contribution provides a complementary perspective, exploring those instances in which not alignment is at stake but rather the contrary: the introduction of new spatial lexical material by addressees.

2 Empirical study

We designed an empirical study to investigate the dynamics of dialogic interaction in a joint spatial task. A scenario was chosen in which it was likely that participants would spontaneously use spatial terms in a variety of ways (as explained above). Pairs of participants were confronted with the task of furnishing a dolls' house. This situation can be characterized as a referential communication task (similar to many earlier studies, e.g., Brennan & Clark, 1996; Brown-Schmidt et al., 2005) combined with joint spatial action (cf. Rickheit & Wachsmuth, 2006). While this corpus is still being prepared for several purposes, we focus here on a subset of data investigated as to the addressee's contribution as just motivated.

2.1 Method and Procedure

For this task, two sets of dolls' house furniture together with two open wooden dolls' houses were used. One of the houses was fully furnished (see Figure 1 below), while the other was empty, with the furniture positioned randomly beside the house. The participants were placed facing each other, but separated by a screen. One of them (henceforth called *matcher*) was placed in front of the empty dolls' house, the other one (henceforth called *director*) in front of the furnished one. Now the director was asked to describe the positions of the furniture in their house in such a way that the matcher could furnish the empty one in exactly the same way. They were encouraged to talk to each other and ask clarification questions, and they were told that the results would be photographed afterwards.

The dialogues (covering between 30 and 90 minutes each) were recorded and transcribed. For present purposes we analyze an extract of the collected data as follows. We focus on the first 50 utterances (segmented according to turn-taking shifts as well as content-related criteria) of 11 different same-sex dyads (3 of which male) of students between 17 and 24 years of age. This way we address a manageable proportion (3.968 words in

total) of the dialogic data that allows for a rough assessment of relative frequencies across a range of participants, while still allowing for a fairly exhaustive qualitative coverage.

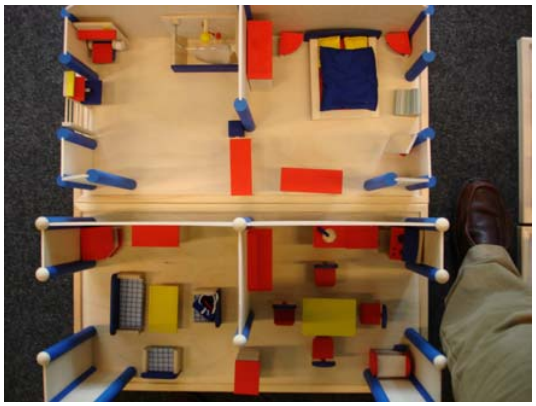


Figure 1. Dolls' house arrangement. During the experiment the dolls' house was arranged with two floors on top of each other and a roof.

2.2 Data Annotation

Prior to the qualitative analysis, we developed three simple (i.e., fairly well definable) annotation steps in order to assess the quantitative relationship of the phenomenon we are interested in with other kinds of dialogue contributions by the matcher.

First, we investigated the *lexical material* contributed by the matcher. We classified all utterances as ACKNOWLEDGEMENTS that do not contain any lexical material other than (typically backgrounded) acknowledgements of the previous instruction (Clark 1996:231; Carletta et al. 1997), expressed by the German equivalents of "yes", "okay", and affirmative feedback signals (*uhuh*). For the remaining utterances, we investigated the extent to which new content is contributed. There are many different ways of distinguishing between *given* and *new* in discourse (e.g., Prince, 1981; Halliday & Matthiessen, 1999). Aiming at the development of operationalizable criteria, we determined for each utterance whether it consists only of lexical material present in the previous discourse context, or whether it introduces new lexical material with respect to the current discourse topic. This way we avoided relying on the subjective interpretation of possible inferences from the earlier discourse. In fact, it is precisely by analyzing the new lexical items that we can gain insights about how inferences are made by the matcher.

Second, we determined whether or not each matcher utterance contained a spatial term, since we are interested in the usage of spatial language. Spatial terms here include various morphological and syntactic forms expressing spatial relationships of any kind, e.g., *left, in front, frontal, middle, to, at, through, out, in, where, here, parallel, there.* From this analysis we extracted those utterances by the matcher that involve the contribution of new spatial content. In our scenario, reaching the discourse goal consists of three steps:

1. *Identifying* an object (out of the range of objects that still need to be placed)
2. *Locating* the object's position in the dolls' house
3. *Orienting* the object in the correct direction in the dolls' house.

The data were also annotated with respect to each of these discourse topics. These steps of analysis were done by two different coders independently, with overlaps for substantial portions of the data and identical annotation results for more than 90%.

3 Results

3.1 Distribution of matcher's utterances

Of a total of 238 utterances by matchers, 98 (41.18%) contained new lexical material, and 114 (47.90%) were ACKNOWLEDGEMENTS without no new content (which are not analyzed further here). Thus, 10.92% of matcher utterances that are not ACKNOWLEDGEMENTS repeat previous lexical material. They can typically be interpreted as REQUESTS FOR EXPANSION (Clark & Wilkes-Gibbs, 1986:22), as in the following example (1):

director: äh die Toilette is äh parallel zur Dusche praktisch an die Hinterwand gestellt. kannst du dir das vorstellen?
[uh the toilet is uh parallel to the shower standing virtually at the back wall. can you imagine that?]
matcher: parallel zur Dusche [parallel to the shower]
director: [provides further information]

82 utterances (34.45% of the total of 238) contained one or more spatial terms; all of these are not classified as ACKNOWLEDGEMENTS. 30 (12.61%) concerned the *identification* of objects, 11 of which contain spatial terms. 62 (26.05%) concern the *location* of objects, 60 of which contain spatial terms; and 16 (6.72%) concern the *orientation* of objects (in 5 cases together with loca-

tion-related content), 12 of which use spatial terms. The remaining 21 utterances concerned other topics; 2 of these contained spatial terms. Thus, spatial terms were mostly used to express *location* or *orientation*. Here are some examples:

Example (2) *Identification:*
dir: da kommt dieser Herd dran,
[there the stove is attached]
match: ein Herd. der mit dem Abzug oben?
[a stove. the one with the hood on top?]

Example (3) *Location:*
dir: ähm dann steht im rechten Zimmer an der Wand das große Bett. [uhm then in the room on right there is the big bed at the wall.]
match: hinten an der Wand? [at the back at the wall?]

Example (4) *Orientation:*
dir: so dass der Kreis so äh zum Bett zeigt.
[so that the circle points uh to the bed]
match: zum Bett? [to the bed?]

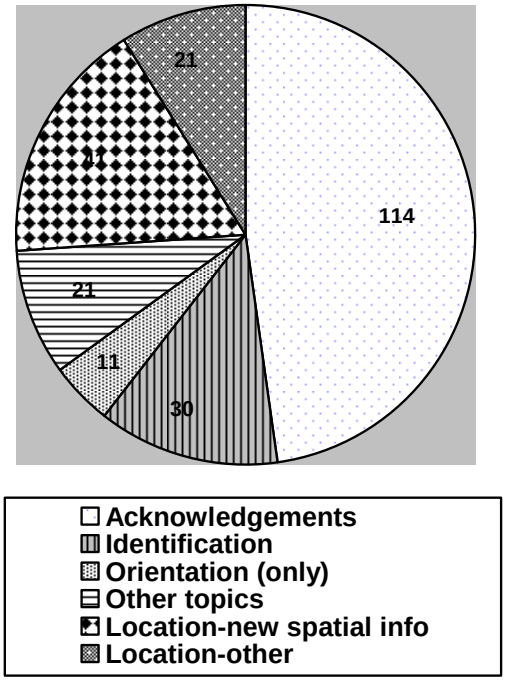


Table 1. Distribution of matcher's utterances

Of the 82 utterances containing a spatial term, 64 contained new lexical material, which did not concern the spatial term in only 8 of the cases. Thus, the matcher regularly contributed *new spatial content* to the dialogue, sometimes for purposes of *identification* of objects, sometimes requesting information about the *orientation* about the object to be placed. In as many as 41 cases (17.23% of all 238 matcher utterances), however, new spatial content was used for *location*-related utterances. In the following subsection we take a closer look at these instances, investigating how the matcher may contribute spatial content to the placement of objects. Table 1 summarizes the categories of matcher's utterances as described so far.

3.2 Negotiation of spatial object location

As Tversky (1999) and others observed, speakers often mix and change perspectives on a spatial scene. In fact, agreeing on a shared perspective poses the most prominent problem in much spatial dialogue research (e.g., Schober, 1993). In our scenario, the director and matcher both have their own dolls' house in front of them so that they share perspective functionally; therefore this kind of conflict should not arise.[1] Nevertheless, there are many ways of conceiving of – and describing – a spatial situation (Tenbrink, 2007); new conceptual perspectives may be added to the information available so far.

Spatial location descriptions consist of three main elements (e.g., Bateman et al., 2007): one (or more) spatial term(s), the *locatum* (the object currently described and – in our scenario – to be placed), and a *relatum* (another object or entity that the locatum is spatially related to) which may remain implicit. Our criterion for identifying new spatial content is based only on the spatial term. To get a clearer idea of how new spatial content is presented we categorized our utterances according to whether the locatum and the relatum, or both, are also new, and develop on this basis a first classification of spatial content suggested by matchers.

All elements new. Utterances that contain new spatial terms, a new locatum, and a new relatum can be said to introduce a completely new spatial description. We identified only three such instances in our data, one of them is example (5):

[1] There are, in fact, a few instances in the data reflecting that the participants did not always realize that perspective was actually shared, as in "also links von meiner Seite aus oder links von deiner Seite aus?" [that is, left from my side or left from your side?] asked by the matcher; this is then clarified by the director: "es steht ja auch vor Dir das Ding das Haus." [it stands in front of you as well you know, the thing the house.]

match: wir sind noch links ne?
[we are still on the left side, right?]

Such instances can be interpreted as clarifying a global aspect of the current situation, removing uncertainty based on the complexity of the task.

Locatum new. There were no instances in which the locatum was new but not the relatum, which would mean that a new object was described in relation to another object that had just been used to describe the position of a different object.

Relatum new. In 16 instances in our data (39.02% of the 41 location-related utterances containing new spatial content), the relatum was new but the locatum was not. Thus, the object currently in focus was described in relation to a different object than the one that the director related it to. In the following example (6), the matcher shows considerable initiative by first summarizing a previous (complex) spatial description by the director (not represented here), and then offering a new description (for the same object location) in addition, marking this explicitly by "also" (that is):

match: und der kommt direkt daneben.
[and this one is put directly beside it]
dir: ja an die Wand ran [yes, at the wall]
match: also hinten links von links von dem Spülbecken.
[that is, at the back left of left of the sink]
dir: ja aber an die Wand so ran an die,
[yes but at the wall at the]

Notice also how the director repeats her own description "an die Wand ran". Apparently these two interlocutors have different conceptions of the scene and wish these to be clarified or confirmed before moving on. Crucially, for the matcher the object location becomes clearer when seen in relation to another object in addition to the one the director chose to relate it to. This possibility arises because of the fact that the matcher has already placed several objects, so that the visual scene offers more than one basis for spatial descriptions. Sometimes the matcher's suggestion of an alternative relatum is used to disambiguate the director's description, as in the following example (7):

match: neben das Klo an die Wand? oder an die andere Wand, [beside the toilet at the wall? or at the other wall]
dir: an die andere Wand. [at the other wall]

In example (8), the matcher's suggestion of a new relatum clarifies a misunderstanding, highlighting an underdeterminacy in the director's instruction:

dir: das Waschbecken stellst Du jetzt so dass das ähm in die Lücke ja okay, [now you place the sink so that it fits into the gap yes okay]
match: ja? zwischen Toilette und Dusche?
[yes? between the toilet and the shower?]
dir: nee nee nee. [no no no.]
match: okay. [okay.]
dir: ähm (…) da fehlt doch so'n Stück der Wand ne?
[uhm (…) there is you know a piece of the wall missing, okay?]

Here, the director offered a "gap" as a relatum, actually not an object but rather a kind of non-entity defined by further entities that are left implicit. The matcher identifies a different interpretation than the one intended and clarifies this by explicitly mentioning the relata she is thinking of (toilet and shower). This induces the director to make the intended underlying relatum (the basis for the non-entity) explicit, namely, the wall.

Another possibility is to offer a first spatial description for the object currently in focus, as in (9):

dir: dann is' das nächste Ding du hast ähm
[then is the next thing that you have uhm]
match: noch immer im selben Raum?
[still in the same room?]
dir: genau. [exactly.]

In this example, the matcher's suggestion remains on a fairly high level of granularity – the relatum "room" together with the spatial term "in" leaves much room for interpretation. The suggestion is based on expectations from the previous discourse context, which the matcher wishes to confirm.

Only spatial term new. In 21 of our 41 cases (51.22%), neither the locatum nor the relatum is new, so that only the spatial term is changed. Typically in these cases, the matcher has detected a spatial ambiguity or underspecification in the director's utterances, which is clarified by changes concerning the spatial term. In five cases in our data (clearly identifiable by the use of "or"), their reaction is to make the options explicit and request a choice, as in the following example (10):

dir: ja genau. stell's an die Wand (…)
[yes, exactly, put it at the wall (…)]
match: frontal , frontal an die Wand,
[frontally, frontally at the wall]
dir: ja genau frontal (...) [yes, exactly, frontally]
match: links oder recht [left or right]
dir: ähm äh, rechts [uhm, uh, right]

Here there are two specific possible positions (left or right) to choose from. In example (11), the problem seems to consist of the area being too large which has so far been determined, so the matcher wishes to clarify the precise position of the object in this area, again by making the options explicit:

dir: an der hinteren Wand dran so dass ähm ja die Füße quasi zu dir zeigen. [at the back wall so that uhm yes the feet point to you so to speak]
match: hinten links oder rechts oder in der Mitte?
[at the back left or right or in the middle?]
dir: ach so genau in der Mitte.
[I see, exactly, in the middle.]

However, not all of the utterances in this category are formulated in this *multiple-choice* fashion. Sometimes the matcher simply adds further (spatial) aspects to the previous description, as in (12):

dir: ähm dann steht im rechten Zimmer an der Wand das große Bett. [uhm then there is the big bed in the room on the right at the wall]
match: hinten an der Wand? [in the back[2] at the wall?]

In other cases, the matcher simply re-formulates the spatial description so that the spatial relationship between locatum and relatum is highlighted in a different way, as in the following example (13):

dir: die steht da so [it stands there in such a way]
match: die passt da so rein?
[it fits in there in such a way?]

In sum. To sum up the results of this subsection, the matcher's contributions of new spatial content in order to locate an object's position may fulfill the following functions:

- to clarify a global aspect of the current situation (using a completely new spatial description)
- to (further) specify an object's position by relating it to an(other) object already placed
- to (further) specify an object's position by suggesting a different or additional spatial term to describe the spatial relationship (in more detail)
- to disambiguate an ambiguous description by explicitly mentioning options.

4 Discussion

How does the addressee (or matcher) contribute to the negotiation of object placement in joint action? In the present study we investigated the matcher's utterances with respect to the extent to which they introduced new lexical material. About half of the matcher's utterances did not contain any new lexical content; these were typically either acknowledgements or requests for expansion. Half of those utterances that did contain new lexical material concerned either the *identification* (prior to its placement) or the *orientation* of an object (after the location has been identified). Only a relatively small number of matcher utterances concerned orientation; this is somewhat surprising given that theoretically objects could be placed in many different orientations. However, in practice the participants may have assumed a standard orientation of the objects according to their expectations as to how dolls' houses should be furnished; and in fact, our arrangements in the present study did not depart from such standard expectations.

Our main interest, however, concerned the other half of those matchers' utterances that contained new lexical material, namely, those negotiating the *location* of objects. Here we determined more closely which part of the utterance was new: the spatial term, the locatum, the relatum, or any combination of these. It turned out that most utterances fell into either one of two major categories, both of which concern the suggestion of a new conceptual perspective on the current spatial scene. Matchers regularly attempted to further specify an object's position by either modifying the spatial term used to describe the position, or by relating the object to another entity that had already been placed.

Why is such an additional specification necessary, and how does the matcher succeed in suggesting spatial content in spite of the fact that only the director has precise knowledge about how the object should be placed? The spatial situation in our scenario – positioning many different objects in a fairly complex array – clearly poses a number of problems such as ambiguity, underspecification, and vagueness. Rather than passively wait for, or simply request, further information in such indeterminate cases, the matchers actively collaborated in identifying the intended location of an object in a range of ways, based on their *assumptions* about probable object locations. These may be derived from various sources, such as the actual spatial situation that is visually accessible to the participants, the previous discourse context, and default assumptions about typical arrangements of objects

[2] As Carroll (1993) points out, German speakers sometimes partition the visual field into regions; more distant positions are then referred to as "back".

in dolls' houses. The consistent setting used here is supportive of this process, as opposed to the Map Task (Anderson et al., 1991), for instance, which uses diverging maps as basis for communication, necessitating additional negotiation processes not inherent to the task itself. Paralleling the seminal findings by Clark & Wilkes-Gibbs (1986), thus, not only *referring* is a collaborative process, but also spatial locations are negotiated jointly, drawing on a well-established set of expectations and a broad range of available conceptual perspectives on the scene.

5 Dialogue structure

How do our findings on dialogic contributions by the matcher relate to previous findings on dialogue structure? There are a number of candidates for dialogue schemes that support categorizing our data in terms of dialogue structure. While they were developed in different contexts and for various purposes, some of the proposed categories match quite straightforwardly to our data, such as ACKNOWLEDGEMENT and REQUEST FOR EXPANSION (cf. Section 3.1 above). In Carletta et al. (1997)'s *move coding scheme* the QUERY-W move matches the disambiguation questions found in our data. The CHECK move "requests the partner to confirm information that the speaker has some reason to believe, but is not entirely sure about." (Carletta et al., 1997:3), which is close to the idea of making concrete suggestions for grounding, although they are in our data typically not represented as requests for confirmation. Clark & Wilkes-Gibbs (1986, pp. 22-24) suggest EXPANSIONS which contain a request for confirmation (of the expansion) while basically accepting the description so far, and REPLACEMENTS which reject the previous description and offer a new one. On a general level, approaches to clarification and grounding procedures found in the literature (e.g., Schlangen, 2004; Purver et al., 2003) fit to our data to a certain extent, though they are typically viewed as contributing to a REPAIR of failing communication, which seems to miss the mark in our case.[3] Clark's notion of SECOND-TURN REPAIR (1996:245), for instance, concerns the clarification of a particular aspect of the description. DAMSL (Allen & Core, 1997) does not have a general category of REPAIR but distinguishes between backward- and forward looking functions. The backward-looking tag HOLD is used in cases where the response to the previous utterance (which may be an instruction as in our context) is postponed pending further clarification. Crucially, this does not signal misunderstanding. As a forward-looking tag, the utterances may further be marked as INFO-REQUEST, defined as "an utterance that creates an obligation for the hearer to provide information."

As yet, none of these approaches capture the finer conceptual distinctions reflected by the usage of spatial language that we have pursued. Our aim in the long run is thus to develop operationalizable criteria for a reliable categorization of each utterance, and to account for the various kinds of principles governing dialogue contributions in spatial contexts.

6 Conclusions and Outlook

In this paper we have investigated the collaborative negotiation of spatial object placement in joint action in a novel naturalistic dialogue setting. Our findings show that addressees actively contribute to the dialogue by offering well-informed suggestions based on their expectations concerning how an object should be placed, specifying earlier descriptions further by suggesting a new conceptual perspective on the scene.

Future work using our dolls' house dialogue corpus will address both dialogue partners' choices of spatial language more closely for example with respect to reference frames and alignment processes, and we will investigate the degree to which features of the scenario influence the addressees' expectations, as reflected in their reactions. In a second line of research, we pursue in our project the modelling of dialogue structure within DAMSL.

Acknowledgements

We wish to thank our collaborators in project I5-[DiaSpace] for extensive discussions on the present topic, and our students for assistance in the experiments and transcription. Funding by the DFG (SFB/TR 8 Spatial Cognition) and a HWK fellowship (Hanse Institute for Advanced Studies, Germany) awarded to Coventry are appreciated.

[3] Sack's notion of 'appendor question' as reported by Schegloff (1997:510f.) seems close to the phenomenon we have described for a spatial context; it is categorized by Schegloff as a form of repair initiation.

References

Allen, James and Mark Core. 1997. *Draft of DAMSL: Dialog Act Markup in Several Layers*. Manuscript, http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/.

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech 34*(4), 351-366.

Bateman, John, Thora Tenbrink, and Scott Farrar. 2007. The Role of Conceptual and Linguistic Ontologies in Discourse. *Dialogue Modelling: Computational and Empirical Approaches. Special Issue of Discourse Processes, 44(3)*, 175–213.

Brennan, Susan E. and Herbert H. Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 22(6): 1482-1493.

Brown-Schmidt, S., E. Campana, and M. K. Tanenhaus. 2005. Real-time reference resolution by naïve participants during a task-based unscripted conversation. In J. Trueswell and M. Tanenhaus (eds.), *World-situated language processing: Bridging the language as product and language as action traditions*. MIT Press, Cambridge, MA, 153-171.

Carletta, Jean, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics 23(1)*, 13–32.

Carroll, Mary. 1993. Deictic and intrinsic orientation in spatial descriptions: a comparison between English and German. In J. Altarriba (ed.), *Cognition and Culture.* Elsevier.

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

Clark, Herbert H. and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language 50*, 62–81.

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

Filipi, Anna and Roger Wales. 2004. Perspective-taking and perspective-shifting as socially situated and collaborative actions. *Journal of Pragmatics, Volume 36, Issue 10.* October 2004, Pages 1851-1884.

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

Levinson, Stephen C. 2003. *Space in Language and Cognition*. Cambridge University Press.

Muller, Philippe and Laurent Prévot (in press). Grounding information in route explanation dialogues. In K. Coventry, T. Tenbrink, and J. Bateman (eds.), *Spatial Language and Dialogue*. Oxford University Press.

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

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In P. Cole (ed.), *Syntax and semantics: Vol. 14. Radical Pragmatics*. New York: Academic Press, pp 223-255.

Purver, Matthew, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt (eds.), *Current and New Directions in Discourse and Dialogue*. Dordrecht: Kluwer, pp. 235--255.

Rickheit, Gert and Ipke Wachsmuth (eds.). 2006. *Situated Communication*. Mouton de Gruyter.

Schegloff, Emanuel A. 1997. Practice and Actions: Boundary Cases of Other-Initiated Repair. *Discourse Processes 23*, 499-545.

Schlangen, David. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of SIGdial04*, Boston, USA.

Schober, Michael F. 1993. Spatial Perspective-Taking in Conversation, *Cognition* 47: 1-24.

Schober, Michael F. (in press). Spatial Dialogue between Partners with Mismatched Abilities. In K. Coventry, T. Tenbrink, and J. Bateman (eds.), *Spatial Language and Dialogue.* Oxford University Press.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Tenbrink, Thora. 2007. *Space, time, and the use of language: An investigation of relationships*. Berlin: Mouton de Gruyter.

Tversky, Barbara. 1999. Spatial Perspective in Descriptions. In P. Bloom, M.A. Peterson, L. Nadel & M.F. Garrett (eds.), *Language and Space* (pp. 109-169). Cambridge, MA: MIT Press.

Watson, Matthew E., Martin J. Pickering, and Holly P. Branigan. 2004. Alignment of Reference Frames in Dialogue. *Proceedings of Cogsci 2004*, Chicago.

van der Zee, Emile and Jon Slack (eds.), 2003. *Representing Direction in Language and Space*. Oxford: Oxford University Press.

Author Index