Evaluating Description and Reference Strategies in a Cooperative Human-Robot Dialogue System

Mary Ellen Foster[1,2] **Manuel Giuliani**[1] **Amy Isard**[2] **Colin Matheson**[2]
Jon Oberlander[2] **Alois Knoll**[1]
[1]Informatik VI: Robotics and Embedded Systems, Technische Universität München
[2]Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh

Abstract

We present a human-robot dialogue system that enables a robot to work together with a human user to build wooden construction toys. We then describe a study which assessed the responses of naïve users to output that varied along two dimensions: the method of describing an assembly plan (*pre-order* or *post-order*), and the method of referring to objects in the world (*basic* and *full*). Varying both of these factors produced significant results: subjects using the system that employed a pre-order description strategy asked for instructions to be repeated significantly less often than those who experienced the post-order strategy, while the subjects who heard references generated by the full reference strategy judged the robot's instructions to be significantly more understandable than did those who heard the output of the basic strategy.

1 Introduction

Numerous interactive systems have addressed 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; examples of such systems include the Bielefeld situated artificial communicator [Knoll *et al.*, 1997], the Leonardo robot developed at MIT [Breazeal *et al.*, 2004] and the NASA Peer-to-Peer human-robot system [Fong *et al.*, 2005].

When humans carry out this type of joint action with one another, they are able to employ a wide range of verbal and non-verbal cues to coordinate their actions; indeed, natural-language (or multimodal) dialogue can itself be seen as a form of joint action [Clark, 1996]. If a robot is to cooperate naturally with a human partner, it must be able to understand and produce the same sort of communicative cues. If the communicative signals are "wrong"—even very subtly—it can have an impact on the usability of a robot system. For example, [Goetz *et al.*, 2003] found that people complied more with a robot whose demeanour matched the seriousness of the task. It has also recently been demonstrated [Huber *et al.*, 2008] that, when a robot arm hands pieces to a human recipient, users react significantly more quickly if the arm uses biologically-inspired motions than if it uses motions created by standard motion-planning algorithms.

In this paper, we describe a user evaluation of a human-robot dialogue system that is designed to enable a humanoid robot to cooperate with a human partner on building wooden construction toys. In the evaluation, we experimentally vary two aspects of the output generated by the system: the way that it describes assembly plans to the user, and the way that it refers to objects in the world. We then measure the impact of varying each of these features on the users' objective success at working with the system, as well as on their subjective impressions of the interaction.

2 Background

The two aspects of the output that were manipulated in this experiment—describing domain plans in dialogue and referring to objects in a multimodal context—are both research areas that have been widely studied in the natural-language community. This section presents an overview of existing approaches to each of these tasks.

2.1 Domain plans and dialogue plans

As part of developing a task-based dialogue system, a critical aspect of the design process is determining the relationship between *domain plans*—that is, plans for achieving task-related goals—and *dialogue plans*—plans for achieving communicative goals. In a task-based interaction, there is a tight relationship between the two types of plans: for example, the system must be able to discuss goals and plan steps, to recognise user actions and integrate them into the plan, and to monitor the execution of the plan as it proceeds.

While it is generally agreed that these two types of plans are tightly related, there is no consensus on exactly how the plans should be linked to one another in practice. One general model for this type of task-based dialogue is the *collaborative problem-solving* (CPS) model of dialogue [Blaylock and Allen, 2005]. In collaborative problem solving, multiple agents jointly select and pursue goals in three interleaved phases: selecting the goals to address, choosing procedures for achieving the goals, and executing the selected procedures. This model was used as the basis for the SAMMIE

in-car dialogue system [Becker *et al.*, 2006], while the COLLAGEN system [Rich *et al.*, 2001] takes a similar view of dialogue and task goals. [Moore, 1995] describes how text structure and domain plan structure can be closely related, while more recent work on automatic tutoring in the LeActiveMath project [Callaway *et al.*, 2006] makes similar assumptions about the interaction between the domain level and dialogue level representations.

2.2 Generation of referring expressions

When agents—human or artificial—work together on a task involving manipulating objects, an important communicative function is indicating to a conversational partner which of a set of available domain entities should be used. In the natural-language generation (NLG) community, this is a core task called *generation of referring expressions*; that is, selecting an expression to identify an entity from a set of entities that can be referred to, in a context available to both the speaker and the hearer. Generating referring expressions, linguistic or multimodal, is one of the classic tasks in NLG, and a number of algorithms have been proposed.

The classic algorithm in reference generation—and the one on which most subsequent implementations are based—is the *incremental algorithm* of [Dale and Reiter, 1995], which selects a set of attributes of a target object to single it out from a set of distractor objects. Attributes are selected repeatedly until only the target object remains in the distractor set. This algorithm ignores much of the context that is available: it aims to generate only initial mentions consisting of noun phrases with articles and modifiers. Several extensions to the incremental algorithm have been proposed to deal with the fact that, in practice, the speaker and the hearer quite often have more context in common. These extensions add notions such as visual and discourse salience [Kelleher and Kruijff, 2006] and the ability to produce multimodal references including actions such as pointing [van der Sluis, 2005; Kranstedt and Wachsmuth, 2005].

[Foster *et al.*, 2008a] noted another type of multimodal reference which is particularly useful in embodied, task-based contexts: *haptic-ostensive* reference, in which an object is referred to as it is being manipulated by the speaker. Manipulating an object, which must be done in any case as part of the task, also makes an object more salient and therefore affords linguistic references that indicate the increased accessibility of the referent. This type of reference is similar to the *placing-for* actions noted by [Clark, 1996].

3 Human-robot dialogue in JAST

The experiment described in this paper makes use of the JAST human-robot dialogue system [Rickert *et al.*, 2007], which supports multimodal human-robot collaboration on a joint construction task. The user and the robot work together to 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 manipulator arms with grippers, mounted in a position to resemble human arms, and an animatronic talking head [van Breemen, 2005] capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The system is able to interact in either English or German.

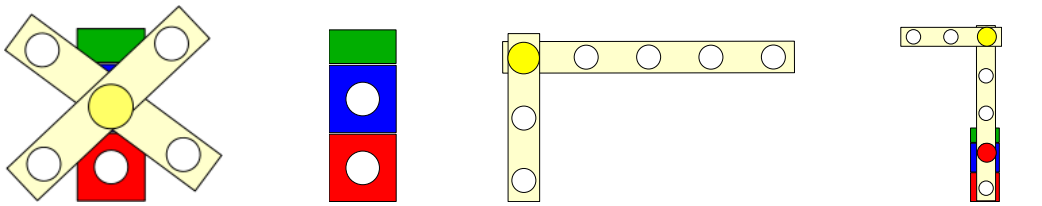
Figure 1: The JAST dialogue robot

The robot is able to manipulate objects in the workspace and to perform simple assembly tasks. In the system used in the current study, the robot instructs the user on building a particular compound object, explaining the necessary assembly steps and retrieving pieces as required, with the user performing the actual assembly actions. The workspace is divided into two areas—one belonging to the robot and one to the human—to make joint action necessary for task success.

Messages on all of the input channels (speech, object recognition, and gesture recognition) are processed and combined by a multimodal fusion component [Giuliani and Knoll, 2008], which sends unified hypotheses to the dialogue manager. The dialogue manager is based on the TrindiKit dialogue management toolkit [Larsson and Traum, 2000], which implements the well-known *information-state based* approach to dialogue management.

As the system works through an assembly plan with the user, the dialogue manager follows one of two strategies for describing the plan. It may use a *pre-order* strategy, in which the structure of the plan is described before moving to specific assembly actions, or it may use a *post-order* strategy, in which it proceeds directly to describing the concrete assembly actions. Figure 2 shows excerpts from sample dialogues using each of these strategies; note that we adopt the same close relationship between domain plans and discourse plans described by [Moore, 1995]. More details on the dialogue manager and on the description strategies are given in [Foster and Matheson, 2008].

Once the dialogue manager has selected a response to the user's input, it sends a high-level specification of the desired



(a) Windmill (b) Snowman (c) L Shape (d) Railway signal

Figure 3: Target objects

Pre-order description strategy, basic reference strategy

SYSTEM First we will build a windmill. Okay?

USER Okay.

SYSTEM To make a windmill, we must make a snowman.

SYSTEM *[picking up and holding out red cube]* To make a snowman, insert the green bolt through the end of the red cube and screw it into the blue cube.

USER *[takes cube, performs action]* Okay.

Post-order description strategy, full reference strategy

SYSTEM First we will build a windmill. Okay?

USER Okay.

SYSTEM *[picking up and holding out red cube]* Insert the green bolt through the end of this red cube and screw it into the blue cube.

USER *[takes cube, performs action]* Okay.

SYSTEM Well done. You have made a snowman.

Figure 2: Sample dialogue excerpts

response to the output planner, which in turn sends commands to produce appropriate output on each of the individual channels to meet the specification: linguistic content (including appropriate multimodal referring expressions), facial expressions and gaze behaviours of the talking head, and actions of the robot manipulators. Two strategies were implemented for generating references to objects in the world: a *basic* version that uses only the basic incremental algorithm to select properties [Dale and Reiter, 1995], and a *full* version that also includes aspects of haptic-ostensive reference [Foster *et al.*, 2008a]. The main distinction between the output of the two versions is that the basic system says 'the red cube' while handing a cube to the user, while the full system instead says 'this red cube' (as in the underlined sentences in Figure 2).

Once the system has described a plan step, the user responds, using a combination of the available input modalities. The user's contribution is processed by the input modules and the fusion component, a new hypothesis is sent to the dialogue manager, and the interaction continues until the target object has been fully assembled.

4 Experiment design

The JAST dialogue system described in the preceding section was evaluated through a user study in which subjects interacted with the complete system; all interactions were in German. Using a between-subjects design, this study compared all of the combinations of the two description strategies with the two reference strategies, measuring the quality of the resulting dialogue, the users' success at building the required objects and at learning the names of new objects, and the users' subjective reactions to the system.

4.1 Subjects

43 subjects (27 male) took part in this experiment; the results of an additional subject were discarded due to technical problems with the system. The mean age of the subjects was 24.5, with a minimum of 14 and a maximum of 55. Of the subjects who indicated an area of study, the two most common areas were Informatics (12 subjects) and Mathematics (10). On a scale of 1–5, subjects gave a mean assessment of their knowledge of computers at 3.4, of speech-recognition systems at 2.3, and of human-robot systems at 2.0. Subjects were compensated for their participation in the experiment.

4.2 Scenario

Each subject built the same three objects in collaboration with the JAST system, always in the same order. The first target object was a *windmill* (Figure 3a), which has a subcomponent called a *snowman* (Figure 3b). After the windmill had been completed, the system then described how to build an *L shape* (Figure 3c). Finally, the robot instructed the user on building a *railway signal* (Figure 3d), which combines an L shape with a snowman.

Before the system explained each target object, the experimenter first configured the workspace with exactly the pieces required to build it. The pieces were always distributed across the two work areas in the same way to ensure that the robot would always hand over the same pieces to each subject. For the windmill, the robot handed over one of the cubes and one of the slats; for the L shape, it handed over both of the required slats; while for the railway signal, it handed over both cubes and both slats.

For objects requiring more than one assembly operation (i.e., all but the L shape), the system gave names to all of the intermediate components as they were built. For example, the windmill was always built by first making a snowman and then attaching the slats to the front, as in the dialogues in

	Pre-order	Post-order
Basic	11	11
Full	11	10

Table 1: Distribution of subjects across the conditions

Figure 2. When the railway signal was being built, the system always asked the user if they remembered how to build a snowman and an L shape. If they did not remember, the robot explained again; if they did remember, the robot simply asked them to build another one using the pieces on the table.

4.3 Independent variables

In this study, we manipulated two independent variables, description strategy and reference strategy, each with two different levels. As explained in Section 3—and as illustrated in Figure 2—the two possible description strategies were *pre-order* and *post-order*, while the two possible reference strategies were *basic* and *full*. Users were assigned to conditions using a between-subjects design, so that each subject interacted with the system using a single combination of description strategy and reference strategy throughout. Subjects were assigned to each combination of factors in turn, following a Latin-square design. As shown in Table 1, 10 subjects interacted with the system that combined the post-order description strategy with the full reference strategy, while each of the other combinations was used by 11 subjects.

4.4 Dependent variables

We measured a wide range of dependent values in this study: objective measures based on the logs and recordings of the interactions, as well as subjective measures based on the users' ratings of their experience. The objective metrics fell into the following three classes, based on those used by [Walker *et al.*, 1997]: *dialogue efficiency*, *dialogue quality*, and *task success*. The subjective metrics were of two types: users rated their emotional state before and after the interaction, and also answered a standard user-satisfaction questionnaire. The full set of dependent variables is given in [Foster *et al.*, 2009].

4.5 Hypotheses

There is no evidence from the existing studies of task-based dialogue to prefer either of the plan-description strategies over the other. Also, previous evaluations of automatically-generated referring expressions (e.g., [Belz and Gatt, 2008]) have found little relationship between the 'naturalness' (i.e., human-likeness) of the references and any task-based measures. We therefore made no specific prediction about the effect of either manipulation on the results of this study.

5 Results

In this study, we gathered a wide range of dependent measures. For the purposes of this paper, we concentrate on the specific impact of the two independent measures that were manipulated. In [Foster *et al.*, 2009], we describe the results of a detailed examination of the relationship among the various subjective and objective measures, using a PARADISE-style analysis [Walker *et al.*, 1997]. This analysis found that

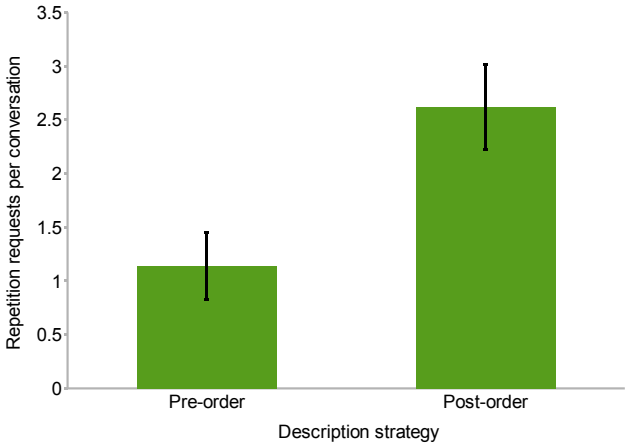


Figure 4: Repetition requests by description strategy

the primary predictors of user satisfaction were the dialogue length, the number of repetition requests, and the users' recall of the system instructions.

To determine the impact of the two independent measures on each of the dependent measures, we performed an ANOVA analysis on each class of dependent measures, using both of the independent measures as categorical predictors—in no case was there a significant interaction between the two factors. We list the primary significant factor for each independent measure below, giving the significance values from the ANOVA analysis. None of the demographic factors (age, gender, area of study, experience with computers) affected any of the results presented here.

5.1 Description strategy

The primary difference between the two description strategies (pre-order vs. post-order) was found on one of the dialogue-quality measures: the rate at which subjects asked the system to repeat itself during an interaction. As shown in Figure 4, subjects in the pre-order condition asked for instructions to be repeated an average of 1.14 times over the course of an interaction, while subjects who used the post-order version of the system asked for repetition 2.62 times on average—that is, more than twice as frequently. The ANOVA analysis indicated that the difference between the two means is significant: $F_{1,39} = 8.28, p = 0.0065$.

5.2 Reference strategy

The choice of referring-expression strategy had no significant effect on any of the objective measures. However, this factor did have an impact on the responses on a set of items from the subjective user-satisfaction questionnaire which specifically addressed the understandability of the robot's instructions. The relevant items are shown in Figure 5. The responses of subjects to these three items was different across the two groups: subjects using the system which employed full referring expressions tended to give higher scores on the first question and lower scores on the second and third, while the responses of subjects using the system with basic referring expressions showed the opposite pattern. The

- Es war einfach den Anweisungen des Roboters zu folgen
  It was easy to follow the robot's instructions
- Der Roboter gab zu viele Anweisungen auf einmal
  The robot gave too many instructions at once
- Die Anweisungen des Roboters waren zu ausführlich
  The robot's instructions were too detailed

Figure 5: Questionnaire items addressing the understandability of the robot's instructions

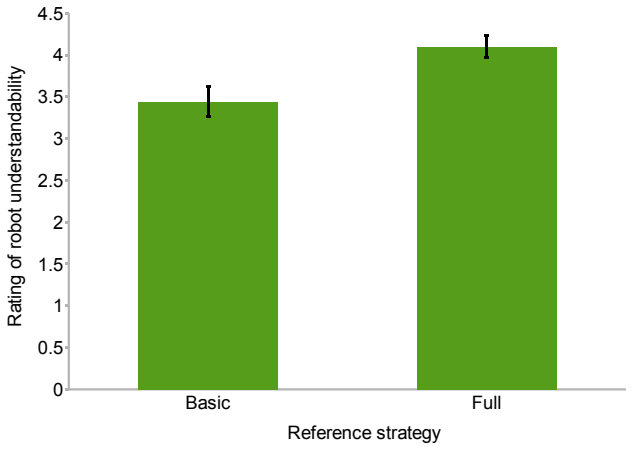


Figure 6: Understandability rating by reference strategy

mean perceived understandability—i.e., the mean of the responses on these three items, using an inverse scale for the latter two—was 3.44 (out of 5) for the system with basic references and 4.10 for the system with full references; these results are shown in Figure 6. The ANOVA analysis indicated that the difference between the two means is significant: $F_{1,39} = 8.32, p = 0.0064$.

6 Discussion

Both of the experimental manipulations had an effect on users' ability to understand the instructions presented by the system, but each at a different level. The choice of description strategy primarily affected how often users asked for instructions to be repeated, but had no effect on their subjective opinions. On the other hand, the choice of reference strategy had no effect on the number of repetition requests, but significantly affect users' subjective impressions of the instructions. Neither of the manipulations had a significant effect on any of the other dependent measures, subjective or objective.

The difference in level is likely due to the different effects that the two manipulations had on the generated output: as shown in Figure 2, using a different description strategy changes the whole progress of the dialogue, while using a different reference strategy makes a more subtle, purely lexical change within a system turn. It appears that, when the system used the less-favoured description strategy, the subjects immediately noticed if they did not understand the instructions; however, the need to ask for repetition did not greatly affect their opinion of the instructions after the fact. On the other hand, the less-preferred reference strategy did not prevent the subjects from understanding the instructions, but does appear to have had an overall effect on their opinions of the system.

7 Conclusions and Future Work

We have reported the results of a user evaluation that was carried out on a human-robot dialogue system. In the study, the robot explained to naïve users how to assemble a collection of objects using a wooden toy construction set. The two independent variables in the study were the robot's strategy for describing the assembly plan to the user and its strategy for generating referring expressions. The two assembly description strategies included a pre-order version in which the robot first labelled an assembly and then explained how to build it, and a post-order version in which the robot gave the instructions first and labelled the assembly afterwards. The two strategies for generating referring expressions included a basic version that used the incremental algorithm of [Dale and Reiter, 1995] and a full version that added aspects of haptic-ostensive reference as described by [Foster *et al.*, 2008a].

The results of the study show that users significantly benefit from the pre-order dialogue strategy over the post-order strategy, which can be seen from the reduced number of times they had to ask the robot to repeat the instructions—this result extends previous findings on the relationship between domain plans and discourse plans and also suggests a preferred strategy for any system that must generate step-by-step instructions. The other significant finding from the study is that users found it easier to follow the robot's instructions when the robot used the full referring expression strategy in comparison to the basic strategy, which supports the current efforts in the natural-language generation community to devise more sophisticated reference-generation algorithms.

For the next version of the JAST system, we plan to support a more sophisticated collaborative assembly scenario in which the user and the robot both know the assembly plan. In such a scenario, the emphasis will shift from explaining the plan to coordinating the actions of the two participants, and a decreased use of verbal communication is expected. Therefore, the non-verbal communication skills of the robot must be improved in order to anticipate and monitor the user's actions and intentions. To achieve these monitoring skills, advanced input modules are necessary, which for example track the user's head orientation to determine the focus of attention, or classify the user's gestures in order to infer their next actions. For error monitoring and goal inference on the robot side, are integrating a non-verbal goal inference component based on dynamic fields [Erlhagen *et al.*, 2007; Foster *et al.*, 2008b]. This will enable the system to monitor the user's action and report any errors to the dialogue manager so that it can determine the correct response, and should also increase the system's ability to respond proactively to the user's needs without needing to be asked. We intend to evaluate this integrated system in an experiment similar to that that described here in order to measure the impact of the added non-verbal goal-inference facilities on the interactions.

Acknowledgements

This research was supported by the European Commission through the JAST project (FP6-003747-IP), `http://www.euprojects-jast.net/`. Thanks to Pawel Dacka for his help in carrying out the experiment.

References

[Becker *et al.*, 2006] T Becker, N Blaylock, C Gerstenberger, I Kruijff-Korbayová, A Korthauer, M Pinkal, M Pitz, P Poller, and J Schehl. Natural and intuitive multimodal dialogue for in-car applications: The SAMMIE system. In *Proceedings of PAIS 2006*, 2006.

[Belz and Gatt, 2008] Anja Belz and Albert Gatt. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL 2008*, 2008.

[Blaylock and Allen, 2005] N Blaylock and J Allen. A collaborative problem-solving model of dialogue. In *Proceedings of SIGdial 2005*, 2005.

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

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

[Clark, 1996] H H Clark. *Using Language*. Cambridge University Press, 1996.

[Dale and Reiter, 1995] R Dale and E Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

[Erlhagen *et al.*, 2007] W Erlhagen, A Mukovskiy, F Chersi, and E Bicho. On the development of intention understanding for joint action tasks. In *Proceedings of ICDL 2007*, 2007.

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

[Foster and Matheson, 2008] M E Foster and C Matheson. Following assembly plans in cooperative, task-based human-robot dialogue. In *Proceedings of Londial 2008*, 2008.

[Foster *et al.*, 2008a] M E Foster, E G Bard, R L Hill, M Guhe, J Oberlander, and A Knoll. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of HRI 2008*, 2008.

[Foster *et al.*, 2008b] M E Foster, M Giuliani, T Müller, M Rickert, A Knoll, W Erlhagen, E Bicho, N Hipólito, and L Louro. Combining goal inference and natural-language dialogue for human-robot joint action. In *Proceedings of the CIMA workshop at ECAI 2008*, 2008.

[Foster *et al.*, 2009] M E Foster, M Giuliani, and A Knoll. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of ACL 2009*, 2009.

[Giuliani and Knoll, 2008] M Giuliani and A Knoll. MultiML: A general-purpose representation language for multimodal human utterances. In *Proceedings of ICMI 2008*, 2008.

[Goetz *et al.*, 2003] J Goetz, S Kiesler, and A Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of RO-MAN 2003*, 2003.

[Huber *et al.*, 2008] M Huber, M Rickert, A Knoll, T Brandt, and S Glasauer. Human-robot interaction in handing-over tasks. In *Proceedings of RO-MAN 2008*, 2008.

[Kelleher and Kruijff, 2006] J D Kelleher and G-J M Kruijff. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of COLING-ACL 2006*, 2006.

[Knoll *et al.*, 1997] A Knoll, B Hildenbrandt, and J Zhang. Instructing cooperating assembly robots through situated dialogues in natural language. In *Proceedings of ICRA 1997*, 1997.

[Kranstedt and Wachsmuth, 2005] A Kranstedt and I Wachsmuth. Incremental generation of multimodal deixis referring to objects. In *Proceedings of ENLG 2005*, 2005.

[Larsson and Traum, 2000] S Larsson and D Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3&4):323–340, September 2000.

[Moore, 1995] J D Moore. The role of plans in discourse generation, 1995. `http://www.pitt.edu/~coconut/discourse-plans.ps`.

[Rich *et al.*, 2001] C Rich, C L Sidner, and N Lesh. COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine*, 22(4):15–25, 2001.

[Rickert *et al.*, 2007] M Rickert, M E Foster, M Giuliani, T By, G Panin, and A Knoll. Integrating language, vision and action for human robot dialog systems. In *Proceedings of HCI International 2007*, 2007.

[van Breemen, 2005] A J N van Breemen. iCat: Experimenting with animabotics. In *Proceedings of AISB 2005 Creative Robotics Symposium*, 2005.

[van der Sluis, 2005] I F van der Sluis. *Multimodal Reference: Studies in Automatic Generation of Multimodal Referring Expressions*. PhD thesis, University of Tilburg, 2005.

[Walker *et al.*, 1997] M A Walker, D J Litman, C A Kamm, and A Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of ACL/EACL 1997*, 1997.