Targeted Help for Spoken Dialogue Systems: intelligent feedback improves naive users' performance

Beth Ann Hockey
Research Institute for Advanced
Computer Science (RIACS),
NASA Ames Research Center,
Moffet Field, CA 94035
bahockey@riacs.edu

Oliver Lemon
School of Informatics,
University of Edinburgh,
2 Buccleugh Place
Edinburgh EH8 9LW, UK
olemon@inf.ed.ac.uk

Ellen Campana
Department of Brain
and Cognitive Sciences
University of Rochester
Rochester, NY 14627
ecampana@bcs.rochester.edu

Laura Hiatt
Center for the Study of Language
and Information (CSLI)
Stanford University
210 Panama St,
Stanford, CA 94305
lahiatt@stanford.edu

Gregory Aist
RIACS
NASA Ames Research Center,
Moffet Field, CA 94035
aist@riacs.edu

James Hieronymus
RIACS
NASA Ames Research Center,
Moffet Field, CA 94035
jimh@riacs.edu

Alexander Gruenstein
BeVocal, Inc.
685 Clyde Avenue
Mountain View, CA 94043
agruenstein@bevocal.com

John Dowding
RIACS
NASA Ames Research Center,
Moffet Field, CA 94035
jdowding@riacs.edu

Abstract

We present experimental evidence that providing naive users of a spoken dialogue system with immediate help messages related to their out-of-coverage utterances improves their success in using the system. A grammar-based recognizer and a Statistical Language Model (SLM) recognizer are run simultaneously. If the grammar-based recognizer suceeds, the less accurate SLM recognizer hypothesis is not used. When the grammar-based recognizer fails and the SLM recognizer produces a recognition hypothesis, this result is used by the Targeted Help agent to give the user feedback on what was recognized, a diagnosis of what was problematic about the utterance, and a related in-coverage example. The in-coverage example is intended to encourage alignment between user inputs and the language model of the system. We report on controlled experiments on a spoken dialogue system for command and control of a simulated robotic helicopter.

1 Introduction

Targeted Help makes use of user utterances that are out-of-coverage of the main dialogue system recognizer to provide the user with immediate feedback, tailored to what the user said, for cases in which the system was not able to understand their utterance. These messages can be much more informative than responding to the user with some variant of "Sorry I didn't understand", which is the behaviour of most current mixed initiative dialogue systems. Providing relevant help messages is a non-trivial problem with mixed initiative systems. There is a much wider range of utterances that the user could sensibly say to a mixed initiative system at any give point in a dialogue. In addition since the system must determine rather than dictate the dialogue state there is uncertainty about the context in which help needs to be given. Our Targeted Help approach is aimed at addressing this problem using information that can reasonably be extracted from imperfect input.

To implement Targeted Help we use two recognizers: the Primary Recognizer is constructed with grammar-based language model and the Secondary Recognizer used by the Targeted Help module is constructed with a Statistical Language Model (SLM). As part of a spoken dialogue system, grammar based recognizers tuned to a domain perform very well, in fact better than comparable Statistical Language Models (SLMs) for in-coverage utterances (Knight et al., 2001). However, in practice users will sometimes produce utterances that are out of coverage. This is particularly true of non-expert users, who do not understand the limitations and capabilities of the system, and consequently produce a much lower percentage of in-coverage utteraces than expert users. The Targeted Help strategy for achieving good performance with a dialogue system is to use a grammar-based language model and assist users in becoming expert as quickly as possible. This approach takes advantage of the strengths of both types of language models by using the grammar based model for in-coverage utterances and the SLM as part of the Targeted Help system for out-of-coverage utterances.

In this paper we report on controlled experiments, testing the effectiveness of an implementation of Targeted Help in a mixed initiative dialogue system to control a simulated robotic helicopter.

2 System Description

2.1 The WITAS Dialogue System

Targeted Help was deployed and tested as part of the WITAS dialogue system[1], a command and control and mixed-initiative dialogue system for interacting with a simulated robotic helicopter or UAV (Unmanned Aerial Vehicle) (Lemon et al., 2001). The dialogue system is implemented as a suite of agents communicating though the SRI Open Agent Architecture (OAA) (Martin et al., 1998). The agents include: Nuance Communications Recognizer (Nuance, 2002); the Gemini parser and generator (Dowding et al., 1993) (both using a grammar designed for the UAV application); Festival text-to-speech synthesizer (Systems, 2001); a GUI which displays a map of the area of operation and shows the UAV's location; the Dialogue Manager (Lemon et al., 2002); the Robot Control and Report component, which translates commands and queries bi-directionally between the dialogue interface and the UAV. The Dialogue Manager interleaves multiple planning and execution dialogue threads (Lemon et al., 2002).

While the helicopter is airborne, an on-board active vision system will interpret the scene below to interpret ongoing events, which may be reported (via NL generation) to the operator. The robot can carry out various activities such as flying to a location, fighting fires, following a vehicle, and landing. Interaction in WITAS thus involves joint-activities between an autonomous system and a human operator. These are activities which the autonomous system cannot complete alone, but which require some human intervention (e.g. search for a vehicle). These activities are specified by the user during dialogue, or can be initiated by the UAV. In any case, a major component of the dialogue, and a way of maintaining its coherence, is tracking the state of current or planned activities of the robot. This system is sufficiently complex to serve as a good testbed for Targeted Help.

2.2 The Targeted Help Module

The Targeted Help Module is a separate component that can be added to an existing dialogue system with minimal changes to accomodate the specifics of the domain. This modular design makes it quite portable, and a version of this agent is in fact being used in a second command and control dialogue system (Hockey et al., 2002a; Hockey et al., 2002b). It is argued in (Lemon and Cavedon, 2003) that "low-level" processing components such as the Targeted Help module are an important focus for future dialogue system research. Figure 1 shows the structure of the Targeted Help component and its relationship to the rest of the dialogue system.

The goal of the Targeted Help system is to handle utterances that cannot be processed by the

[1] See http://www.ida.liu.se/ext/witas and http://www-csli.stanford.edu/semlab/witas

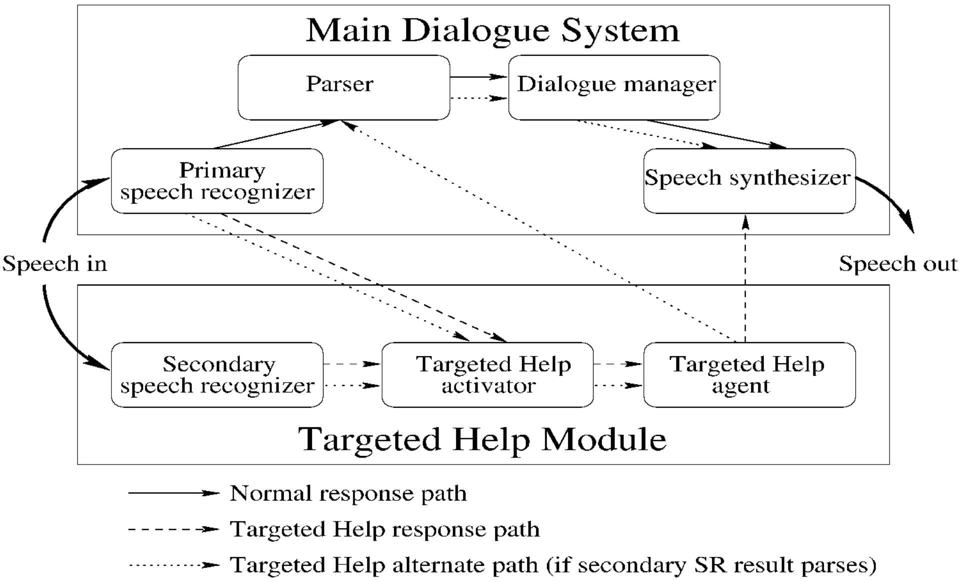

Figure 1: Architecture of Dialogue System with Targeted Help Module

usual components of the dialogue system, and to align the user's inputs with the coverage of the system as much as possible. To perform this function the Targeted Help component must be able to determine which utterances to handle, and then construct help messages related to those utterances, which are then passed to a speech synthesizer. The module consists of three parts:

- the Secondary Recognizer,
- the Targeted Help Activator,
- the Targeted Help Agent.

The Targeted Help Activator takes input from both the main grammar-based recognizer and the backup category-based SLM recognizer. It uses this input to determine when the Targeted Help component should produce a message. The Activator's behavior is as follows for the four possible combinations of recognizer outcomes:

1. Both recognizers get a recognition hypothesis:
   Targeted Help remains inactive; normal dialogue system processing proceeds
2. Main recognizer gets a recognition hypothesis and secondary recognizer rejects:
   Targeted Help remains inactive; normal dialogue system processing proceeds
3. Main recognizer rejects, secondary recognizer gets a recognition hypothesis and secondary recognizer hypothesis can be parsed (rare):
   normal dialogue system processing continues using the secondary recognizer output
4. Main recognizer rejects, secondary recognizer gets a recognition hypothesis and secondary recognizer hypothesis cannot be parsed :
   Targeted Help is activated
5. Both recognizers reject:
   Targeted Help is not activated, default system failure message is produced

Once Targeted Help is activated, the Targeted Help Agent constructs a message based on the recognition hypothesis from the secondary SLM recognizer. These messages are composed of one or more of the following pieces:

What the system heard: a report of the backup SLM recognition hypothesis.

What the problem was: a description of the problem with the user's utterance (e.g. the system doesn't know a word); and

What you might say instead: A similar in-coverage example.

In constructing both the diagnostic of the problem with the utterance, and the in-coverage example, we are faced with the question of whether the information from the secondary recognizer is sufficient to produce useful help messages. Since this domain is relatively novel, there is not very much data for training the SLM and the performance reflects this. We have designed a rule based system that looks for patterns in the recognition hypothesis that seem to be detected adequately even with incomplete or inaccurate recognition.

Diagnostics are of three major types:

- endpointing errors,
- unknown vocabulary,
- subcategorization mistakes.

We found from an analysis of transcripts that these three types of errors accounted for the majority of failed utterances. Endpointing errors are cases of one or the other end of an utterance being cut off. For example, when the user says "search for the red car" but the system hears "for the red car". We use information from the dialogue system's parsing grammar (which has identical coverage to its speech recognizer) to determine whether the initial word recognized for an utterance is a valid initial word in the grammar. If not, the utterance is diagnosed as a case of the user pressing the push-to-talk button too late and the system reports that to the user.[2] Out-of-vocabulary items that can be identified by Targeted Help are those that are in the SLM's vocabulary but are out of coverage for the grammar based recognizer and so cannot be processed by the dialogue system. For these items Targeted Help produces a message of the form "the system doesn't understand the word X".

Saying "Zoom in on the red car" when the system only has intransitive "zoom in" is an example of a subcategorization error. In these cases the word is in-vocabulary but has been used in a way that is out-of-grammar. This is not simply a deficiency of the grammar. In this case, for example, zooming in on a particular object is not part of the functionality of the system. To diagnose subcategorization errors we consult the recognition/parsing grammar for subcategorization information on in-vocabulary verbs in the secondary recognizer hypothesis, then check what else was recognized to determine if the right arguments are there. For these types of errors the system produces a message such as "the system doesn't understand the word X used with *the red car*". These diagnostics are one substantive difference from the approach used in (Gorrell et al., 2002). The simple classifier approach used in that work to select example sentences would not support these types of diagnostics.

In constructing examples that are similar to the user's utterance one issue is in what sense they should be similar. One aspect we have looked at is using in-coverage words from the user's utterance. It is likely to help naive users learn the coverage of the system if the examples give them valid uses of in-coverage words they produced in their utterance. By using words from the user's utterance the system provides both confirmation that those words are in coverage and an in-coverage pattern to imitate. We believe that this leads to greater linguistic alignment between the user and the system. Another aspect of similarity that we suspect is important is matching the utterance dialogue-move type (e.g. wh-question, yes/no-question, command) otherwise the user is likely to be misled into thinking that a particular type of dialogue-move is impossible in the system.

Looking for in-coverage words is fairly robust. Even when the user produces an out-of-coverage utterance they are likely to produce some in-coverage words. The Targeted Help agent looks for within-domain words in the recognition hypothesis from the secondary SLM recognizer. This gives us a set of target words from which to match the example to the dialogue-move type of the user's utterance: wh-question, yn-question, answer, or command.

Furthermore, for commands (which are a large percentage of the utterances) we use the in-coverage words to produce a targeted in-coverage

[2] while this problem may seem peculiar to the use of push-to-talk, in fact using another approach such as open microphone simply introduces different endpointing (and other) problems. Whatever system is employed, users will still need to learn how it works to perform well with the system.

example that is interpretable by the system. These examples are intended to demonstrate how in-vocabulary words from the backup recognizer hypothesis could be successfully used in communicating with the system. For example, if the user says something like "fly over to the hospital", where "over" is out-of-coverage, and the fallback recognizer detected the words "fly" and "hospital", the Targeted Help agent could provide an in-coverage example like "fly to the hospital". For the other less frequent utterance types we have one in coverage example per type. The system currently uses a look-up table but we hope to incorporate generation work which would support generation of these examples on the fly from a list of in-coverage words (Dowding et al., 2002).

3 Design of Experiments

In order to assess the effectiveness of the targeted help provided by our system, we compared the performance of two groups of users, one that received targeted help, and one that did not. Twenty members of the Stanford University community were randomly assigned to one of the two groups. There were both male and female subjects, the majority of subjects were in their twenties and none of the subjects had prior experience with spoken dialogue systems. The structure of the interaction with the system was the same for both groups. They were given minimal written instruction on how to use the system before the interaction began. They were then asked to use the system to complete five tasks, in which they directed a helicopter to move within a city environment to complete various task oriented goals which were different for four of the five tasks. For each task the goals were given immediately prior to the start of the interaction, in language the system could not process to prevent users from simply reading the goal aloud to the system. A given task ended when one of the following criteria was met:

1. the task was accurately completed and the user indicated to the system that he or she had finished,
2. the user believed that the task was completed and indicated this to the system when in fact the task was not accurately completed, or
3. the user gave up.

The first and last of the sequence of five tasks were the critical trials that were used to assess performance. Both of the tasks had goals of the form "locate an x and then land at the y" The experiment was conducted in a single session. An experimenter was present throughout, but when asked she refused to provide any feedback or hints about how to interact with the system.

As stated above, the critical difference between the two groups of users was the feedback they received during interaction with the system. When the users in the No Help condition produced out-of-coverage utterances the system responded only with a text display of the message "not recognized". In contrast, when users in the Help condition produced out-of-coverage utterances they received in-depth feedback such as: "The system heard *fly between the hospital and the school*, unfortunately it doesn't understand *fly* when used with the words *between the hospital and the school*. You could try saying *fly to the hospital*."

We hypothesized that: 1) providing Targeted Help would improve users' ability to complete tasks (HIGHER TASK COMPLETION); and 2) time to complete tasks would be reduced for users receiving Targeted Help (REDUCED TIME). We also anticipated that both effects would be more marked in the first task than in the fifth task (LARGER EARLY EFFECT).

4 Experimental Results

We found clear evidence that targeted help improves performance in this environment, as measured by both the frequency with which the user simply explicitly gave up on a task, and the time to complete the remaining tasks. In this section we present the statistical analyses of the experiment. For the following analyses two subjects, both in the No Help condition, were excluded from the analyses because they gave up on every task, leaving 9 users in each of the two help conditions. Exceptions are noted.

We begin by examining the percentage of trials in which users explicitly gave up on a task before it was completed. We compared the percentage of trials in which the user clicked the "give up" button in both tasks for users in both help conditions. As predicted, a 1-within (Task), 1-between (Help condition) subjects ANOVA revealed a main effect of the help condition ($F_1(1,16)=6.000$, $p<.05$). Users who received targeted help were less likely to give up than those who did not receive help, particularly during the first task (11% vs. 27%). If we include the two subjects in the No Help condition who gave up on every task the difference is even more striking. For the first task only 11% of the users who received help gave up, compared to 45% of the users who did not receive help. The pattern holds up even if we include the three intervening filler trials along with the experimental trials, as demonstrated by a paired t-test item analysis ($t(4) = 7.330$, $p<.05$). Those who received help were less likely to explicitly give up even on this wider variety of tasks.

We next examine the time it took users to complete the individual tasks. Here it is necessary to be clear about what is meant by "completion." It is more ambiguous than it may seem. Each task had several sub-goals, and it was even difficult to objectively evaluate whether a single sub goal had been met. For instance, the goal of the first task was to find a red car near the warehouse and then land the helicopter. Users tended to indicate that they had finished as soon as they saw the red car, failing to land the helicopter as the instructions specified. Another common source of ambiguity was when the user saw the car on the map but never brought it up in the dialogue, simply landing the helicopter and clicking "finished." The problem with this is that there is no way of knowing whether the user actually saw the car before clicking finish, and there was no explicit record that they were aware of its presence. For all trials the experimenter evaluated the task completion, recording what was done and what was left undone. According to the experimenter, in most cases of potential ambiguity the basic goal was completed. In a few instances, however, the user indicated belief that the task had been completed when it obviously had not. An example of this is the following: The goal specified was to find a red car near the warehouse and then land. The user flew the helicopter to the police station, and then clicked "finished," ending the task. We dealt with the ambiguity problem by analyzing the time to completion data separately according to two different inclusion criteria. In both cases the pattern was the same: Users who received help took less time to complete tasks than those who did not, the first task took longer to complete than the last one, and the difference between the help and no help conditions was more marked on the first task than on the last one.

In the first analysis we included all trials in which the user clicked the "finished" button, regardless of their actual performance. Subjects who failed to complete one of the two critical tasks (tasks 1 and 5) were excluded from the analysis. We used a 1-within (Task), 1-between (Help condition) subjects ANOVA. For task 1, 89% of the trials in the Help condition and 55% of the trials in the No Help were considered "completed." For task 5, 100% of the trials in the Help condition and 80% of the trials in the No Help condition were considered "completed." The analysis revealed a marginally significant main effect of the help condition ($F_1(1,11) = 3.809$, $p<.1$), a main effect of task ($F_{1,11}=62.545$, $p <.001$) and a help condition by task interaction ($F_1(1,11)=10.203$, $p < .05$). The effects were in the predicted direction. Users who received help took less time to complete tasks than those who did not (290.4 seconds vs. 440.6 seconds), the first task took longer to complete than the last one (365.5 seconds vs. 220.4 seconds), and the difference between the help and no help conditions was more marked on the first task than on the last one (150.2 seconds vs. 94 seconds). Figure 2 shows these results.

One criticism of this analysis is that it may include trials in which the task objectives were not accurately completed before the subject clicked "finished". We wished to avoid experimenter subjectivity with respect to task completion, so we conducted another analysis using the strictest inclusion criterion the experimental design allowed. In this analysis we included only those trials in which all task objectives were completed and could be verified using the transcripts. This meant that for all of the trials we included, the goal entity was explicitly mentioned in the dialogue. According to this criterion only 44% of users in the Help condition and 18% of users in the No Help con-

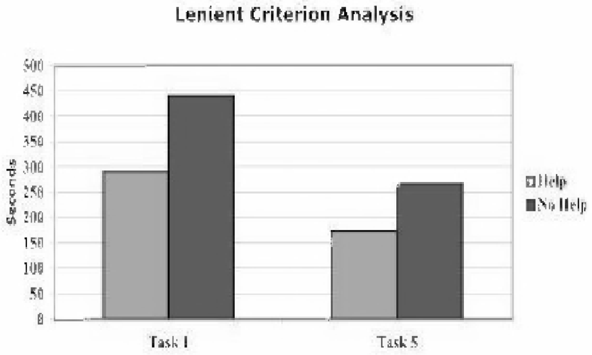

Figure 2: Time to complete task under Lenient Criterion for completion

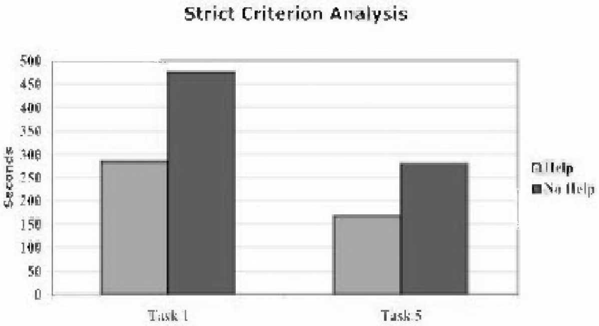

Figure 3: Time to complete task under Strict Criterion for completion

dition completed the first task. Similarly, 89% of users in the Help condition and 40% of users in the No Help condition accurately completed the task. Although this analysis is conducted on sparse data, it provides strong supporting evidence for the data pattern observed in the more lenient analysis.

We examined the time it took to complete tasks according to the strict criterion, excluding all other trials. The ANOVA analysis was identical to the previous one. It, too, revealed a main effect of help condition ($F_1(1,3) = 15.438$, $p<.05$), a main effect of task ($F_{1,3}$=83.512, $p < .01$), and a help condition by task interaction ($F_1(1,3)$=20.335, $p < .05$). Again the effects were in the predicted direction. Users who received help took less time to complete tasks than those who did not (226.2 seconds vs. 377.5 seconds), the first task took longer to complete than the last one (379.9 seconds vs. 223.75), and the difference between the help and no help conditions was more marked on the first task than on the last one (190.4 seconds vs. 112.3 seconds). These results are shown in Figure 3.

5 Conclusions

We have shown that users benefit from having online Targeted Help. Naive users who received Targeted Help messages were less likely to give up and significantly faster to complete tasks than users who did not. Overall, those who did not receive help gave up on 39% of the trials, while those who received our Targeted Help only gave up on 6% of the trials. With respect to time, when we considered all trials in which the user indicated that the goal had been completed (regardless of performance), those users who did not receive our Targeted Help took 53% longer than those who did. Under stricter inclusion criteria, which required the users to explicitly mention the goal and accurately complete the task, the difference was even more pronounced. Those users who did not receive help took 67.0% longer to complete the tasks than those who received our Targeted Help. In both help conditions, performance improved over the course of the experimental session. However, the advantage conferred by help merely diminished and did not disappear during the session.

These findings are remarkable because they demonstrate that it is possible to construct effective Targeted Help messages even from fairly low quality secondary recognition. Moreover, the study suggests that such an approach can improve the speed of training for naive users, and may result in lasting improvements in the quality of their understanding.

6 Future Work

This work suggests many interesting directions for further research. One area of investigation is the contribution of various factors in the effectiveness of the Targeted Help message for example:

- What benefit is due to the online nature of the help?
- What benefit is due to the information content?

- What is the relative contribution of the various parts of the Targeted Help message to the improvement in user performance.
  - Is the diagnostic alone more or less effective than the example alone?
  - How much does getting the back up recognizer hypothesis help the user?
  - What is the most effective combination of these components?

Another interesting direction is to look at effectiveness across different types of applications. The fact that we found positive results in this domain and that (Gorrell et al., 2002) also found a variant of Targeted Help useful on a quite different domain suggests that the approach could be generally useful for a variety of types of dialogue systems. We are currently looking at porting our Targeted Help agent to additional domains.

Acknowledgements

This work was partially funded by the Wallenberg Foundation's WITAS project, Linköping University, Sweden and partially funded through RIACS under NASA Cooperative Agreement Number NCC 2-1006.

References

J. Dowding, M. Gawron, D. Appelt, L. Cherny, R. Moore, and D. Moran. 1993. Gemini: A natural language system for spoken language understanding. In *Proceedings of the Thirty-First Annual Meeting of the Association for Computational Linguistics.*

J. Dowding, G. Aist, B.A. Hockey, and E. O. Bratt. 2002. Generating canonical examples using candidate words. In *(under submission).*

G. Gorrell, I. Lewin, and M. Rayner. 2002. Adding intelligent help to mixed-initiative spoken dialogue systems. In *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP)*, Denver, CO.

B.A. Hockey, G. Aist, J. Dowding, and J. Hieronymus. 2002a. Targeted help and dialogue about plans. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (demo track)*, Philadelphia, PA.

B.A. Hockey, G. Aist, J. Hieronymus, O. Lemon, and J. Dowding. 2002b. Targeted help: Embedded training and methods for evaluation. In *Intelligent Tutoring Systems (ITS) Workshop on Empirical Methods*, San Sebastian, Spain.

S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling, and I. Lewin. 2001. Comparing grammar-based and robust approaches to speech understanding: a case study. In *Proceedings of Eurospeech 2001*, pages 1779–1782, Aalborg, Denmark.

Oliver Lemon and Lawrence Cavedon. 2003. Multilevel architectures for natural-activity-oriented dialogues. In *Proceedings of EACL 2003 workshop on Dialogue Systems: interaction, adaptation and styles of management*, page (in press).

Oliver Lemon, Anne Bracy, Alexander Gruenstein, and Stanley Peters. 2001. Information states in a multimodal dialogue system for human-robot conversation. In Peter Kühnlein, Hans Reiser, and Henk Zeevat, editors, *5th Workshop on Formal Semantics and Pragmatics of Dialogue (Bi-Dialog 2001)*, pages 57 - 67.

Oliver Lemon, Alexander Gruenstein, and Stanley Peters. 2002. Collaborative activities and multitasking in dialogue systems. *Traitement Automatique des Langues (TAL)*, 43(2):131 - 154. Special Issue on Dialogue.

D. Martin, A. Cheyer, and D. Moran. 1998. Building distributed software systems with the open agent architecture. In *Proceedings of the Third International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*, Blackpool, Lancashire, UK.

Nuance, 2002. http://www.nuance.com. As of 1 Feb 2002.

The Festival Speech Synthesis Systems, 2001. http://www.cstr.ed.ac.uk/projects/festival. As of 28 February 2001.