Spontaneous Speech Understanding for Robust Multi-Modal Human-Robot Communication

Sonja Hüwel, Britta Wrede
Faculty of Technology, Applied Computer Science
Bielefeld University, 33594 Bielefeld, Germany
shuewel,bwrede@techfak.uni-bielefeld.de

Abstract

This paper presents a speech understanding component for enabling robust situated human-robot communication. The aim is to gain semantic interpretations of utterances that serve as a basis for multi-modal dialog management also in cases where the recognized word-stream is not grammatically correct. For the understanding process, we designed semantic processable units, which are adapted to the domain of situated communication. Our framework supports the specific characteristics of spontaneous speech used in combination with gestures in a real world scenario. It also provides information about the dialog acts. Finally, we present a processing mechanism using these concept structures to generate the most likely semantic interpretation of the utterances and to evaluate the interpretation with respect to semantic coherence.

1 Introduction

Over the past years interest in mobile robot applications has increased. One aim is to allow for intuitive interaction with a personal robot which is based on the idea that people want to communicate in a natural way (Breazeal et al., 2004)(Dautenhahn, 2004). Although often people use speech as the main modality, they tend to revert to additional modalities such as gestures and mimics in face-to-face situations. Also, they refer to objects in the physical environment. Furthermore, speech, gestures and information of the environment are used in combination in instructions for the robot. When participants perceive a shared environment and act in it we call this communication "situated" (Milde et al., 1997). In addition to these features that are characteristic for situated communication, situated dialog systems have to deal with several problems caused by spontaneous speech phenomena like ellipses, indirect speech acts or incomplete sentences. Large pauses or breaks occur inside an utterance and people tend to correct themselves. Utterances often do not follow a standard grammar as written text.

Service robots have not only to be able to cope with this special kind of communication but they also have to cope with noise that is produced by their own actuators or the environment. Speech recognition in such scenarios is a complex and difficult task, leading to severe degradations of the recognition performance. The goal of this paper is to present a framework for human-robot interaction (HRI) that enables robust interpretation of utterances under the specific conditions in HRI.

2 Related Work

Some of the most explored speech processing systems are telephone-based information systems. Their design rather differs from that of situated HRI. They are uni-modal so that every information has to be gathered from speech. However, speech input is different as users utter longer phrases which are generally grammatically correct. These systems are often based on a large corpus and can therefore be well trained to perform satisfactory speech recognition results. A prominent example for this is the telephone based weather forecast information service JUPITER (Zue et al., 2000).

[1]This work has been supported by the European Union within the 'Cognitive Robot Companion' (COGNIRON) project (FP6-IST-002020) and by the German Research Foundation within the Graduate Program 'Task Oriented Communication'.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 391–398,
Sydney, July 2006.

Over the past years interest increased in mobile robot applications where the challenges are even more complex. While many of these problems (person tracking, attention, path finding) are already in the focus of research, robust speech understanding has not yet been extensively explored in the context of HRI. Moreover, interpretation of situated dialogs in combination with additional knowledge sources is rarely considered. Recent projects with related scope are the mobile robots CARL (Lopes et al., 2005) and ALBERT (Rogalla et al., 2002), and the robotic chandelier Elvis (Juster and Roy, 2004). The main task of the robot CARL is robust language understanding in context of knowledge acquisition and management. It combines deep and shallow parsing to achieve robustness. ALBERT is designed to understand speech commands in combination with gestures and object detection with the task to handle dishes. The home lighting robot Elvis gets instructions about lighting preferences of a user via speech and gestural input. The robot itself has a fixed position but the user may walk around in the entire room. It uses keyword spotting to analyze the semantic content of speech. As speech recognition in such robot scenarios is a complex and difficult task, in these systems the speech understanding analysis is constrained to a small set of commands and not oriented towards spontaneous speech. However, deep speech understanding is necessary for more complex human robot interaction.

There is only little research in semantic speech analysis of spontaneous speech. A widely used approach of interpreting sentences is the idea of case grammar (Bruce, 1975). Each verb has a set of named slots, that can be filled by other slots, typically nouns. Syntactic case information of words inside a sentence marks the semantic roles and thus, the corresponding slots can be filled. Another approach of processing spontaneous speech by using semantic information for the Air Travel Information Service (ATIS) task is implemented in the Phoenix system (Ward, 1994). Slots in frames represent the basic semantic entities known to the system. A parser using semantic grammars maps input onto these frame representations. The idea of our approach is similar to that of the Phoenix system, in that we also use semantic entities for extracting information. Much effort has been made in the field of parsing strategies combined with semantic information. These systems support preferably task oriented dialog systems, e.g., the ATIS task as in (Popescu et al., 2004) and (Milward, 2000), or virtual world scenarios (Gorniak and Roy, 2005), which do not have to deal with uncertain visual input. The aim of the FrameNet project (Fillmore and Baker, 2001) is to create a lexicon resource for English, where every entry receives a semantic frame description.

In contrast to other presented approaches we focus on deep semantic analysis of situated spontaneous speech.Written language applications have the advantage to be trainable on large corpora, which is not the case for situated speech based applications. And furthermore, interpretation of situated speech depends on environmental information. Utterances in this context are normally less complex, still our approach is based on a lexicon that allows a broad variety of utterances. It also takes speech recognition problems into account by ignoring non-consistent word hypotheses and scoring interpretations according to their semantic completeness. By adding pragmatic information, natural dialog processing is facilitated.

3 Situated Dialog Corpus

With our robot BIRON we want to improve social and functional behavior by enabling the system to carry out a more sophisticated dialog for handling instructions. One scenario is a home-tour where a user is supposed to show the robot around the home. Another scenario is a plant-watering task, where the robot is instructed to water different plants. There is only little research on multimodal HRI with speech-based robots. A study how users interact with mobile office robots is reported in (Hüttenrauch et al., 2003). However, in this evaluation, the integration of different modalities was not analyzed explicitly. But even though the subjects were not allowed to use speech and gestures in combination, the results support that people tended to communicate in a multi-modal way, nevertheless.

To receive more detailed information about the instructions that users are likely to give to an assistant in home or office we simulated this scenario and recorded 14 dialogs from German native speakers. Their task was to instruct the robot to water plants. Since our focus in this stage of the development of our system lies on the situatedness of the conversation, the robot was simply replaced by a human pretending to be a robot. The subjects

were asked to act as if it would be a robot. As proposed in (Lauriar et al., 2001), a preliminary user study is necessary to reduce the number of repair dialogs between user and system, such as queries. The corpus provides data necessary for the design of the dialog components for multi-modal interaction. We also determined the lexicon and obtained the SSUs that describe the scene and tasks for the robot.

The recorded dialogs feature the specific nature of dialog situations in multi-modal communication situations. The analysis of the corpus is presented in more detail in (Hüwel and Kummert, 2004). It confirms that spontaneously spoken utterances seldom respect the standard grammar and structure of written sentences. People tend to use short phrases or single words. Large pauses often occur during an utterance or the utterance is incomplete. More interestingly, the multi-modal data shows that 13 out of 14 persons used pointing gestures in the dialogs to refer to objects. Such utterances cannot be interpreted without additional information of the scene. For example, an utterance such as "this one" is used with a pointing gesture to an object in the environment. We realize, of course, that for more realistic behavior towards a robot a real experiment has to be performed. However this time- and resource-efficient procedure allowed us to build a system capable of facilitating situated communication with a robot. The implemented system has been evaluated with a real robot (see section 7). In the prior version we used German as language, now the dialog system has adapted to English.

4 The Robot Assistant BIRON

The aim of our project is to enable intuitive interaction between a human and a mobile robot. The basis for this project is the robot system BIRON (et. al, 2004). The robot is able to visually track persons and to detect and localize sound sources.

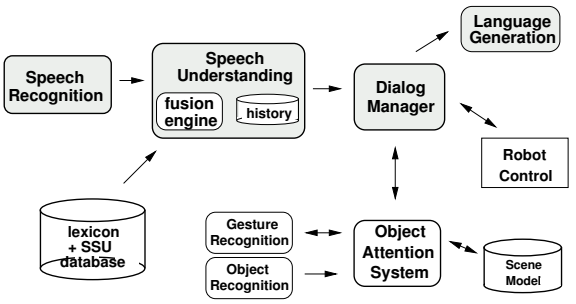

Figure 1: Overview of the BIRON dialog system architecture

The robot expresses its focus of attention by turning the camera into the direction of the person currently speaking. From the orientation of the person's head it is deduced whether the speaker addresses the robot or not. The main modality of the robot system is speech but the system can also detect gestures and objects. Figure 1 gives an overview of the architecture of BIRON's multi-modal interaction system. For the communication between these modules we use an XML based communication framework (Fritsch et al., 2005). In the following we will briefly outline the interacting modules of the entire dialog system with the speech understanding component.

Speech recognition: If the user addresses BIRON by looking in its direction and starting to speak, the speech recognition system starts to analyze the speech data. This means that once the attention system has detected that the user is probably addressing the robot it will route the speech signal to the speech recognizer. The end of the utterance is detected by a voice activation detector. Since both components can produce errors the speech signal sent to the recognizer may contain wrong or truncated parts of speech. The speech recognition itself is performed with an incremental speaker-independent system (Wachsmuth et al., 1998), based on Hidden Markov Models. It combines statistical and declarative language models to compute the most likely word chain.

Dialog manager: The dialog management serves as the interface between speech analysis and the robot control system. It also generates answers for the user. Thus, the speech analysis system transforms utterances with respect to gestural and scene information, such as pointing gestures or objects in the environment, into instructions for the robot. The dialog manager in our application is agent-based and enables a multi-modal, mixed ini-

tiative interaction style (Li et al., 2005). It is based on semantic entities which reflect the information the user uttered as well as discourse information based on speech-acts. The dialog system classifies this input into different categories as e.g., instruction, query or social interaction. For this purpose we use discourse segments proposed by Grosz and Sidner (Grosz and Sidner, 1986) to describe the kind of utterances during the interaction. Then the dialog manager can react appropriately if it knows whether the user asked a question or instructed the robot. As gesture and object detection in our scenario is not very reliable and time-consuming, the system needs verbal hints of scene information such as pointing gestures or object descriptions to gather information of the gesture detection and object attention system.

5 Situated Concept Representations

Based on the situated conversational data, we designed "situated semantic units" (SSUs) which are suitable for fast and automatic speech understanding. These SSUs basically establish a network of strong (mandatory) and weak (optional) relations of sematic concepts which represent world and discourse knowledge. They also provide ontological information and additional structures for the integration of other modalities. Our structures are inspired by the idea of frames which provide semantic relations between parts of sentences (Fillmore, 1976).

Till now, about 1300 lexical entries are stored in our database that are related to 150 SSUs. Both types are represented in form of XML structures. The lexicon and the concept database are based on our experimental data of situated communication (see section 3) and also on data of a home-tour scenario with a real robot. This data has been annotated by hand with the aim to provide an appropriate foundation for human-robot interaction. It is also planned to integrate more tasks for the robot as, e.g., courier service. This can be done by only adding new lexical entries and corresponding SSUs without spending much time in reorganization. Each lexical entry in our database contains a semantic association to the related SSUs. Therefore, equivalent lexical entries are provided for homonyms as they are associated to different concepts.

In figure 2 the SSU *Showing* has an open link to the SSUs *Actor* and *Object*. Missing links to

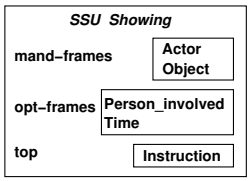

Figure 2: *Schematic SSU "Showing" for utterances like "I show you my poster tomorrow".*

strongly connected SSUs are interpreted as missing information and are thus indicators for the dialog management system to initiate a clarification question or to look for information already stored in the scene model (see fig. 1). The SSUs also have connections to optional arguments, but they are less important for the entire understanding process.

The SSUs also include ontological information, so that the relations between SSUs can be described as general as possible. For example, the SSU *Building_subpart* is a sub-category of *Object*. In our scenario this is important as for example the unit *Building_subpart* related to the concept"wall" has a fixed position and can be used as navigation-support in contrast to other objects. The top-category is stored in the entry *top*, a special item of the SSU. By the use of ontological information, SSUs also differentiate between task and communication related information and thereby support the strategy of the dialog manager to decouple task from communication structure. This is important in order to make the dialog system independent of the task and enable scalable interaction capabilities. For example the SSU *Showing* belongs to the discourse type *Instruction*. Other types important for our domain are *Socialization*, *Description*, *Confirmation*, *Negation*, *Correction*, and *Query*. Further types may be included, if necessary.

In our domain, missing information in an utterance can often be acquired from the scene. For example the utterance "look at this" and a pointing gesture to a table will be merged to the meaning "look at the table". To resolve this meaning, we use hints of co-verbal gestures in the utterance. Words as "this one" or "here" are linked to the SSU *Potential_gesture*, indicating a relation between speech and gesture. The timestamp of the utterance enables temporal alignment of speech and gesture. Since gesture recognition is expensive in computing time and often not well-defined, such linguistic hints can reduce these costs dra-

matically.

The utterance “that” can also represent an anaphora, and is analyzed in both ways, as anaphora and as gesture hint. Only if there is no gesture, the dialog manager will decide that the word probably was used in an anaphoric manner.

Since we focus on spontaneous speech, we cannot rely on the grammar, and therefore the semantic units serve as the connections between the words in an utterance. If there are open connections interpretable as missing information, it can be inferred what is missing and be integrated by the contextual knowledge. This structure makes it easy to merge the constituents of an utterance solely by semantic relations without additional knowledge of the syntactic properties. By this, we lose information that might be necessary in several cases for disambiguation of complex utterances. However, spontaneous speech is hard to parse especially since speech recognition errors often occur on syntactically relevant morphemes. We therefore neglect the cases which tend to occur very rarely in HRI scenarios.

6 Semantic Processing

In order to generate a semantic interpretation of an utterance, we use a special mechanism, which unifies words of an utterance into a single structure. The system also considers the ontological information of the SSUs to generate the most likely interpretation of the utterance. For this purpose, the mechanism first associates lexical entries of all words in the utterance with the corresponding SSUs. Then the system tries to link all SSUs together into one connected uniform. Some SSUs provide open links to other SSUs, which can be filled by semantic related SSUs. The SSU *Beside* for example provides an open link to *Object*. This SSU can be linked to all *Object* entities and to all subtypes of *Object*. Thus, an utterance as ”next to the door” can be linked together to form a single structure (see fig. 3). The SSUs which possess open links are central for this mechanism, they represent roots for parts of utterances. However, these units can be connected by other roots, likewise to generate a tree representing semantic relations inside an utterance.

The fusion mechanism computes in its best case in linear time and in worst case in square time. A scoring function underlies the mechanism: the more words can be combined, the better is the rating. The system finally chooses the structure with the highest score. Thus, it is possible to handle semantic variations of an utterance in parallel, such as homonyms. Additionally, the rating is helpful to decide whether the speech recognition result is reliable or not. In this case, the dialog manager can ask the user for clarification. In the next version we will use a more elaborate evaluation technique to yield better results such as rating the amount of concept-relations and missing relations, distinguish between important and optional relations, and prefer relations to words nearby.

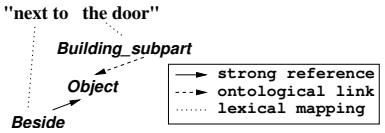

Figure 3: *Simplified parse tree example* .

A converter forwards the result of the mechanism as an XML-structure to the dialog manager. A segment of the result for the dialog manager is presented in Figure 4. With the category-descriptions the dialog-module can react fast on the user’s utterance without any further calculation. It uses them to create inquiries to the user or to send a command to the robot control system, such as “look for a gesture”, “look for a blue object”, or “follow person”. If the interpreted utterance does not fit to any category it gets the value *fragment*. These utterances are currently interpreted in the same way as partial understandings and the dialog manager asks the user to provide more meaningful information.

Figure 1 illustrates the entire architecture of the speech understanding system and its interfaces to other modules. The SSUs and the lexicon are stored in an external XML-databases. As the speech understanding module starts, it first reads these databases and converts them into internal data-structures stored in a fast accessible hash table. As soon as the module receives results from speech recognition, it starts to merge. The mechanism also uses a *history*, where former parts of utterances are stored and which are also integrated in the fusing mechanism. The speech understanding system then converts the best scored result into a semantic XML-structure (see Figure 4) for the dialog manager.

```
<metaInfo>
<time>1125573609635</time>
<status>full</status>
</metaInfo>
<semanticInfo>
<u>what can you do</u>
<category>query</category>
<content>
  <unit = Question_action>
     <name>what</name>
     <unit = Action>
        <name>do</name>
        <unit = Ability>
           <name>can</name>
           <unit = Proxy>
              <name>you</name>
        ...

<u>this is a green cup</u>
<category>description</category>
<content>
  <unit = Existence>
     <name>is</name>
     <unit = Object_kitchen>
        <name>cup</name>
        <unit = Potential_gesture>
           <name>this</name>
        </unit>
        <unit = Color>
           <name>green</name>
        </unit>
      ...
```

Figure 4: Two segments of the speech understanding results for the utterances *"what can you do" and "this is a green cup"*.

6.1 Situated Speech Processing

Our approach has various advantages dealing with spontaneous speech. Double uttered words as in the utterance "look - look here" are ignored in our approach. The system still can interprete the utterance, then only one word is linked to the other words. Corrections inside an utterance as "the left em right cube" are handled similar. The system generates two interpretations of the utterance, the one containing left the other right. The system chooses the last one, since we assume that corrections occur later in time and therefore more to the right. The system deals with pauses inside utterances by integrating former parts of utterances stored in the *history*. The mechanism also processes incomplete or syntactic incorrect utterances. To prevent sending wrong interpretations to the dialog-manager the scoring function rates the quality of the interpretation as described above. In our system we also use scene information to evaluate the entire correctness so that we do not only have to rely on the speech input. In case of doubt the dialog-manager requests to the user.

For future work it is planned to integrate additional information sources, e.g., inquiries of the dialog manager to the user. The module will also store these information in the *history* which will be used for anaphora resolution and can also be used to verify the output of the speech recognition.

```
User1: Robot look - do you see?
       This - is a cow. Funny.
       Do you like it? ...
User2: Look here robot - a cup.
       Look here a - a keyboard.
       Let's try that one. ...
User3: Can you walk in this room?
       Sorry, can you repeat your answer?
       How fast can you move? ...
```

Figure 5: Excerptions of the utterances during the experiment setting.

7 Evaluation

For the evaluation of the entire robot system BIRON we recruited 14 naive user between 12 and 37 years with the goal to test the intuitiveness and the robustness of all system modules as well as its performance. Therefore, in the first of two runs the users were asked to familiarize themselves with the robot without any further information of the system. In the second run the users were given more information about technical details of BIRON (such as its limited vocabulary). We observed similar effects as described in section 2. In average, one utterance contained 3.23 words indicating that the users are more likely to utter short phrases. They also tend to pause in the middle of an utterance and they often uttered so called meta-comments such as "that"s fine". In figure 5 some excerptions of the dialogs during the experiment settings are presented.

Thus, not surprisingly the speech recognition error rate in the first run was 60% which decreased in the second run to 42%, with an average of 52%. High error rate seems to be a general problem in settings with spontaneous speech as other systems also observed this problem (see also (Gorniak and Roy, 2005)). But even in such a restricted experiment setting, speech understanding will have to deal with speech recognition error which can never be avoided.

In order to address the two questions of (1) how well our approach of automatic speech understanding (ASU) can deal with automatic speech recognition (ASR) errors and (2) how its performance compares to syntactic analysis, we performed two analyses. In order to answer question (1) we compared the results from the semantic analysis based on the real speech recognition re-

sults with an accuracy of 52% with those based on the really uttered words as transcribed manually, thus simulating a recognition rate of 100%. In total, the semantic speech processing received 1642 utterances from the speech recognition system. From these utterances 418 utterances were randomly chosen for manual transcription and syntactic analysis. All 1642 utterances were processed and performed on a standard PC with an average processing time of 20ms, which fully fulfills the requirements of real-time applications. As shown in Table 1 39% of the results were rated as complete or partial misunderstandings and 61% as correct utterances with full semantic meaning. Only 4% of the utterances which were correctly recognized were misinterpreted or refused by the speech understanding system. Most errors occurred due to missing words in the lexicon.

Thus, the performance of the speech understanding system (ASU) decreases to the same degree as that of the speech recognition system (ASR): with a 50% ASR recognition rate the number of non-interpretable utterances is doubled indicating a linear relationship between ASR and ASU.

For the second question we performed a manual classification of the utterances into syntactically *correct* (and thus parseable by a standard parsing algorithm) and *not-correct*. Utterances following the English standard grammar (e.g. imperative, descriptive, interrogative) or containing a single word or an NP, as to be expected in answers, were classified as correct. Incomplete utterances or utterances with a non-standard structure (as occurred often in the baby-talk style utterances) were rated as not-correct. In detail, 58 utterances were either truncated at the end or beginning due to errors of the attention system, resulting in utterances such as "where is", "can you find", or "is a cube". These utterances also include instances where users interrupted themselves. In 51 utterances we found words missing in our lexicon database. 314 utterances where syntactically correct, whereas in 28 of these utterances a lexicon entry is missing in the system and therefore would lead to a failure of the parsing mechanism. 104 utterances have been classified as syntactically not-correct.

	ASR=100%	ASR=52%
ASU not or part. interpret.	15%	39%
ASU fully interpretable	84%	61%

Table 1: Semantic Processing results based on different word recognition accuracies.

In contrast, the result from our mechanism performed significantly better. Our system was able to interprete 352 utterances and generate a full semantic interpretation, whereas 66 utterances could only be partially interpreted or were marked as not interpretable. 21 interpretations of the utterances were semantically incorrect (labeled from the system wrongly as correct) or were not assigned to the correct speech act, e.g., "okay" was assigned to no speech act (*fragment*) instead to *confirmation*. Missing lexicon entries often lead to partial interpretations (20 times) or sometimes to complete misinterpretations (8 times). But still in many cases the system was able to interprete the utterance correctly (23 times). For example "can you go for a walk with me" was interpreted as "can you go with me" only ignoring the unknown "for a walk".The utterance "can you come closer" was interpreted as a partial understanding "can you come" (ignoring the unknown word "closer"). The results are summarized in Table 2.

As can be seen the semantic error rate with 15% non-interpretable utterances is just half of the syntactic correctness with 31%. This indicates that the semantic analysis can recover about half of the information that would not be recoverable from syntactic analysis.

	ASU		Synt. cor.
not or part. interpret.	15%	not-correct	31%
fully interpret.	84%	correct	68%

Table 2: Comparison of semantic processing result with syntactic correctness based on a 100% word recognition rate.

8 Conclusion and Outlook

In this paper we have presented a new approach of robust speech understanding for mobile robot assistants. It takes into account the special characteristics of situated communication and also the difficulty for the speech recognition to process utterances correctly. We use special concept structures for situated communication combined with an automatic fusion mechanism to generate semantic structures which are necessary for the dialog manager of the robot system in order to respond adequately.

This mechanism combined with the use of our

SSUs has several benefits. First, speech is interpreted even if speech recognition does not always guarantee correct results and speech input is not always grammatically correct. Secondly, the speech understanding component incorporates information about gestures and references to the environment. Furthermore, the mechanism itself is domain-independent. Both, concepts and lexicon can be exchanged in context of a different domain.

This semantic analysis already produces elaborated interpretations of utterances in a fast way and furthermore, helps to improve robustness of the entire speech processing system. Nevertheless, we can improve the system. In our next phase we will use a more elaborate scoring function technique and use the correlations of mandatory and optional links to other concepts to perform a better evaluation and also to help the dialog manager to find clues for missing information both in speech and scene. We will also use the evaluation results to improve the SSUs to get better results for the semantic interpretation.

References

C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda. 2004. Humanoid robots as cooperative partners for people. *Int. Journal of Humanoid Robots*.

B. Bruce. 1975. Case systems for natural language. *Artificial Intelligence*, 6:327–360.

K. Dautenhahn. 2004. Robots we like to live with?! - a developmental perspective on a personalized, lifelong robot companion. In *Proc. Int. Workshop on Robot and Human Interactive Communication (RO-MAN)*.

A. Haasch et. al. 2004. BIRON – The Bielefeld Robot Companion. In E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, editors, *Proc. Int. Workshop on Advances in Service Robotics*, pages 27–32. Fraunhofer IRB Verlag.

C. J. Fillmore and C. F. Baker. 2001. Frame semantics for text understanding. In *Proc. of WordNet and Other Lexical Recources Workshop*. NACCL.

C. J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conf. on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

J. Fritsch, M. Kleinehagenbrock, A. Haasch, S. Wrede, and G. Sagerer. 2005. A flexible infrastructure for the development of a robot companion with extensible HRI-capabilities. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 3419–3425.

P. Gorniak and D. Roy. 2005. Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. In *ICMI*. ACM Press.

B. J. Grosz and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

H. Hüttenrauch, A. Green, K. Severinson-Eklundh, L. Oestreicher, and M. Norman. 2003. Involving users in the design of a mobile office robot. *IEEE Transactions on Systems, Man and Cybernetics, Part C*.

S. Hüwel and F. Kummert. 2004. Interpretation of situated human-robot dialogues. In *Proc. of the 7th Annual CLUK*, pages 120–125.

J. Juster and D. Roy. 2004. Elvis: Situated Speech and Gesture Understanding for a Robotic Chandelier. In *Proc. Int. Conf. Multimodal Interfaces*.

S. Lauriar, G. Bugmann, T. Kyriacou, J. Bos, and E. Klein. 2001. Personal robot training via natural language instructions. *IEEE Intelligent Systems, 16:3*, pages 38–45.

S. Li, A. Haasch, B. Wrede, J. Fritsch, and G. Sagerer. 2005. Human-style interaction with a robot for cooperative learning of scene objects. In *Proc. Int. Conf. on Multimodal Interfaces*.

L. Seabra Lopes, A. Teixeira, M. Quindere, and M. Rodrigues. 2005. From robust spoken language understanding to knowledge aquisition and management. In *EUROSPEECH 2005*.

J. T. Milde, K. Peters, and S. Strippgen. 1997. Situated communication with robots. In *First Int. Workshop on Human-Computer-Conversation*.

D. Milward. 2000. Distributing representation for robust interpretation of dialogue utterances. In *ACL*.

A.-M. Popescu, A. Armanasu, O. Etzioni, D. Ko, and A. Yates. 2004. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In *Proc. of COLING*.

O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, and R. Dillmann. 2002. Using gesture and speech control for commanding a robot assistant. In *Proc. of the 11th IEEE Int. Workshop on Robot and Human interactive Communication*, pages 454–459. ROMAN.

S. Wachsmuth, G. A. Fink, and G. Sagerer. 1998. Integration of parsing and incremental speech recognition. In *EUSIPCO*, volume 1, pages 371–375.

W. Ward. 1994. Extracting Information From Spontaneous Speech. In *ICRA*, pages 83–86. IEEE Press.

V. Zue, S. Seneff, J. Glass, J. Polifronti, C. Pao, T. J. Hazen, and L. Hetherington. 2000. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, pages 100–112, January.