FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D1.2005

Joint RA1 Deliverable

Multi-modal dialogs

Due date of deliverable: 31/12/2005
Actual submission date: 25/01/2006

Start date of project: January 1st, 2004 **Duration:** 48 months

Contributing partners:
UBi, KTH, UH, LAAS

Revision: final
Dissemination Level: PU

Executive Summary

Objectives in RA1 for the second phase included the implementation of a more flexible dialog system as a basis for adaptation (WP1.1), the integration of multi-modal object reference resolution (CF ROR) and modality-independent object representation in cooperation with RA2 (WP1.2) and the annotation and analysis of the human-robot interaction data gathered in the first phase (WP1.3).
In WP1.1 the finite-state-based dialog system was replaced by a more powerful grounding-based dialog system. This system realizes a dialog model that represents the agent-based view of dialog: a dialog is a collaboration between two intelligent agents and grounding regulates this collaboration. The two main contributions of this model are (1) that it is based on Clark's theory on grounding [9] and (2) that it explicitly integrates a multi-modal account of dialog contributions.
In a second line of work (WP1.2) we focused on the resolution of multi-modal object references in the human-robot dialog. The three essential achievements of this work were (1) the design of a mechanism to systematically handle multi-modal input, (2) the interaction with an object attention system developed in RA2 that can focus on objects in the scene based on the description given verbally by the user, and (3) a representation of objects that both the dialog system and the object attention system can handle.
The evaluation of the multi-modal dialog (WP1.3) is crucially dependent on the use of reliable data collection methods. The work on evaluation of dialog has focused on the work with an annotated multi-modal corpus, based on the experiments in Phase 1. In the second phase the development of the corpus with respect to the level and type of annotation has been continued.

Role of multi-modal dialogs in Cogniron

Using language to communicate with others is one of the most important cognitive abilities of humans. Enabling dialog capability is, therefore, essential for a cognitive robot companion which is supposed to demonstrate human-like capabilities. Since a robot is embodied and situated in the real environment the dialog system of a robot has to handle more complex interactions than in human computer interaction. One of the crucial aspects is the handling of multi-modality because in embodied communication human interlocutors make heavy use of gestures and other non-verbal signals and make references to the shared environment. Building a flexible dialog system with the ability to handle multi-modal information and continuously evaluating the system during the different development cycles are therefore the focus of this research activity.

Relation to the Key Experiments

We have been concentrating on the home tour scenario in KE1. In this experiment, the interaction with the user is mainly carried out via the dialog system. The robot acquires information about the home environment via speech and gesture and forwards it to other system modules for further processing. The dialog system is, therefore, the main interaction component in this experiment. In the other two KE's it is also possible to demonstrate the dialog capability of the robot in a meaningful way.

1 Multi-modal dialogs

1.1 Adaptive multi-modal dialog

The main focus of work in this workpackage carried out by UniBi with contributions from KTH and UH with respect to design issues was the implementation of a more flexible dialog framework that enables the integration of multi-modal cues and adaptation to the user. The main characteristics are the grounding-based approach and the multi-modal account of dialog contributions.

1. **Modeling grounding.** Clark [9] proposed the notion of grounding: during a conversation the interlocutors need to coordinate their mental states based on their mutual understanding about the current intentions, goals and tasks. This means, one can only react to one's dialog partner's contribution meaningfully (providing **acceptance**) if she has understood what her partner has said (**presentation**), in other words, if the common ground is available. Our dialog model regulates the dialog initiative distribution and the discourse management based on this idea. We represent interlocutors' contributions as exchanges. They reach the state "grounded" only if the acceptance of the presentation is available which depends on the communication success (e.g., if the speech input is clearly understood) and the robot task execution status. These exchanges are organized in a stack which represents the ungrounded discourse up to the current state. The grounding status of the whole stack is dependent on the status of the individual exchanges and the relations between them. We introduced 4 types of such relations (default, support, correct and delete). Each contribution of the interlocutors (user and robot) is categorized in terms of its role, i.e., if it initiates an exchange of a certain relation to the previous exchange or if it is the acceptance of an existing one. According to this role, either a new exchange is pushed onto the stack or an old one (or a group of old ones) is popped because it reaches the status "grounded". All the popped exchanges are collected in order to record the complete dialog history. We thus model the grounding process using an augmented push-down automaton which exhibits local flexibility in contrast to conventional approaches ([10], [8]). The implemented system enables a mixed-initiative dialog style and can handle complex repair behaviors.

2. **Integrating multi-modality handling ability.** We take into account the complete language production process and extend the conventional representation of a contribution to a two-layered structure which we call *Interaction Unit (IU)*. Each IU has an intention and conversation layer that are responsible for generation of intention (based on the robot control system's messages) and the generation of language (based on the intention) in verbal and nonverbal form. This representation enables the system to systematically study both the verbal and the non-verbal parts of the user input (e.g., speech and gesture) and also to generate multi-modal output (speech and facial expression).

This model allows us to easily implement several important features: Firstly, a mixed-initiative ability of the system which allows a more complex behavior of the dialog by enabling a freer structure of dialog exchanges. This is a consequence of the user studies carried out in the first phase where users experienced difficulties with the restricted dialog structure.
Secondly, the resolution of multi-modal object references which was done in cooperation with RA2 (see WP 1.2 for detail).
Thirdly, an online communication success measurement that is based on the number of support exchanges in the on-going discourse segment. Such a measurement is important for the later integration of adaptation strategies.

And, finally, two dialog modes representing different levels of system initiative. This feature provides an interesting mechanism for studying social aspects of human-robot interaction and the perception of the robot in user studies as shown in a first evaluation where users were asked to rate the robot's personality traits.
The detail of the model, the implemented system and a first evaluation have been documented in [4] (submitted). First results on the perception of personality traits - inspired by the user studies on personality performed at UH - are summarized in the attached paper [5] (submitted). These findings are currently being complemented by a questionnaire study assessing users' expectations of personality traits of robots in different scenarios (household, child care, reception).
In 2004 and 2005 the team at the University of Hertfordshire performed some basic research in the area of interaction styles, personal spaces and social rules in human-robot interaction as part of their work for WP3.1. This work has implications for the interactive robot programs being developed primarily for KE1 in RA1. There have been continuing contacts throughout the reporting period at the various COGNIRON meetings and by email.
The main results from the UH studies have been documented more fully in the associated COGNIRON deliverables in RA 3. In addition, a short paper has been circulated (refer to D 3.6.1.) which contains all the main relevant results in a short summary form which enables easy use by partners of our findings in guiding the development of robot programs and systems which need to incorporate socially acceptable behavior for a given HRI situation or scenario. More formally, [6] provides further in-depth details with other related papers on robot social spaces available (see D3.1.1 for more details).

1.2 Representation and integration of knowledge for an embodied multi-modal dialog system

In this workpackage we followed three essential lines of work: (1) The design of a mechanism to systematically handle multi-modal input, (2) the interaction with an object attention system developed in RA2 that can focus on objects in the scene based on users' specifications given by speech and gestures, and (3) a representation of objects that both the dialog system and the object attention system can handle. Work in this workpackage was mainly performed by UniBi.

1. **The mechanism provided by the dialog system.** The dialog system represents interlocutors' contributions using a two-layered structure, the so called *Interaction Unit (IU)*. When analyzing the user's input, the dialog system first studies the verbal generator on the conversation layer, i.e., the speech input of the user. If there are unclear object references in this generator the dialog system will consult the non-verbal generator. Here, the object attention system will be activated to search for the object in the current scene. If the search was successful the dialog system proceeds to create its own IU. If no object could be identified, the dialog system initiates a new support exchange to ask for further specifications.

2. **The object attention system (OAS).** The OAS was developed within RA2 and integrated with the dialog module within the home tour scenario (KE1). The OAS receives queries from the dialog system based on the object specifications extracted from the user's verbal utterance, such as its name or color. Upon the request from the dialog system the OAS will start the search in the current scene with the help of a gesture recognizer. In case of successful search the OAS assigns an ID to the object and stores the available information about the object (ID, image, name and other attributes) in a memory and informs the dialog system. If the search was not successful, the OAS sends a query to the dialog system to ask for further specifications of the object.

3. **The representation of the object.** The objects found in the scene are stored in a memory that is augmented with a modality converter. This converter is able to select the appropriate format of data according to the module that initiates a query. For example, if the dialog system queries the color of an object, the symbolic name of the color will be returned. However, if the query comes from the OAS the color will be returned in terms of an HSI value (Hue Saturation Intensity). This converter thus enables a shared representation and usage of object information from different modules based on different modalities.

The resolution of object references has been successfully implemented for KE1, the home tour scenario where the user can now teach the robot objects by speech and deictic gestures. The dialog system initiates clarification questions in cases of failure.
A detailed description of this work has been published (in cooperation with RA2) at the ICMI 2005 in [1].

1.3 Evaluation methods

The evaluation of a multi-modal dialog is crucially dependent on the use of reliable data collection methods. KTH's work on evaluation of dialog has focused on the work with an annotated multi-modal corpus, based on experiments in Phase 1.
In the second phase we have continued to develop the corpus with respect to the level and type of annotation. The corpus is primarily concerned with the home tour scenario described in KE1 and is based on the experimental data that has been collected in the first phase of the project. We are using a coding taxonomy for communicative acts that can be viewed as a multi-modal extension of the DAMSL taxonomy [7], with respect to referential (deictic) gestures. In addition to transcriptions of communicative acts, i.e., speech and gesture, we are also providing annotations of the relevant communicative context, e.g., positioning, task information [2] (submitted).
The corpus has been used in the design process by providing data for an initial analysis of miscommunication [3] (submitted). The result of the analysis has been used to inform the design of the natural language user interface that is being developed in WP1.1. The analysis yielded approximately 20 types of miscommunication. Some of the most frequent types are listed below (cf. Table 1). Further analysis of these generated a set of design implications, concerning the quality and type of feedback given by the system.

Miscommunication type	**Design implication**
Users' knowledge of system capabilities	Explicit feedback related to problem is needed
	Tutorial introducing general features to first time users
Problems related to feedback, Non-relevant response	Non-verbal information to give faster and simultaneous feedback on system state
	Feedback on when it is appropriate to talk, i.e., prompting and alignment to user
Ambiguous or underspecified reference	Visual feedback display (screen) or deictic gestures should be displayed by the robot

Table 1: Some of the most frequent types of miscommunication and design implications.

There remain challenges in the dialog, for instance miscommunication directly related to performance of speech recognition components. Speech recognition failures lead to timing errors, when the robot speaks while the user is providing another command. Although speech recognition poses a main challenge for natural language user interface research, we are concentrating our efforts to challenges that are specific to human-robot interaction. Primarily we are focusing on reducing miscommunication related to the users' limited knowledge of system capabilities by exploring recovery strategies and measuring communicative success to perform online adaptation. In a joint discussion between KTH and UniBi about the corpus analysis and the design implications derived from it two main types of adaptation were identified:

1. **User-specific adaptation:** where models are based on user types identified in the corpus (e.g. cooperative vs. non-cooperative, talkative vs. non-talkative). These models may be initialized using information gathered in an introductory tutorial. In the tutorial, basic functions and interaction capabilities of the robot are explained. Throughout the tutorial specific information about the user is acquired, for example concerning interaction preferences.

2. **Generic adaptation:** where adaptation strategies are selected based on an evaluation of the current interaction situation, regardless of the specific user preferences. Generic adaptation models may use task-specific information derived from a planning component, to select an appropriate adaptation strategy; for instance, expectations about the next required action of the robot such as a request to fetch a missing object. Generic adaptation strategies may further be selected based on a measurement of the current interaction quality. These measures will be based on corpus data that provide a wide collection of multi-modal user behavior such as utterances, gestures and gaze direction that can be used to identify symptoms of miscommunication.

Further design implications are related to the situatedness of the interaction and include, for example, specific means to facilitate the reference resolution by enabling the robot to point to referenced objects or locations or to allow for directive steering commands.

2 Future Work

In the next phase our focus is to enhance the adaptivity of the dialog system to enable a more personalized and intelligent dialog with different users. We will build explicit user models and adopt different dialog strategies according to the user type based on the analysis of human-robot interaction data. In the last phase of the project we will extend this adaptivity to situation awareness so that the robot can select its interaction strategies by also taking into account situational changes. The design and evaluation of this new capability will be carried out through collaboration with RA2 for the detection of human activities, RA5 for spatial representation, and RA6 for exploring the integration of dialog and action planning. UH (RA3) and UniBi will explore possible collaborations for new studies to be conducted in 2006 and 2007 in terms of integrating the dialog system in HRI studies at UH.

3 References

3.1 Applicable documents

Published

[1] Shuyin Li, Axel Haasch, Britta Wrede, Jannik Fritsch, and Gerhard Sagerer. Human-style interaction with a robot for cooperative learning of scene objects. In *Proc. Int. Conf. on Multimodal Interfaces*, pages 151–158, Trento, Italy, 2005. ACM Press.

Submitted

[2] Anders Green, Helge Hüttenrauch, Elin Anna Topp, and Kerstin Severinson Eklundh. Developing a contextualized multimodal corpus for human-robot interaction. Submitted.

[3] Anders Green, Britta Wrede, Shuyin Li, and Kerstin Severinson Eklundh. Integrating miscommunication analysis in natural language interface design for a service robot. Submitted.

[4] Shuyin Li, Britta Wrede, and Gerhard Sagerer. An agent-based, multi-modal dialog system for human robot interaction. Submitted.

[5] Britta Wrede, Stephan Buschkämper, and Shuyin Li. Do you like this robot? The role of robot behavior, robot personality and user personality. Submitted.

3.2 Reference documents

Related Project Papers

[6] Michael L Walters, Kerstin Dautenhahn, René te Boekhorst, Kheng Lee Koay, Christina Kaouri, Sarah Woods, Chrystopher Nehaniv, David Lee, and Iain Werry. The influence of subject's personality traits on personal spatial zones in a human-robot interaction experiment. In *Proc. ROMAN*, 2005.

Other References

[7] J. Allen and M. Core. Draft of DAMSL: Dialog act markup in several layers. http://www.cs.rochester.edu/ research/cisd/ resources/damsl/RevisedManual/, 1997.

[8] J. E. Cahn and S. E. Brennan. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[9] H. H. Clark, editor. *Arenas of Language Use*. University of Chicago Press, 1992.

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

4 Annexes

The annex contains a set of papers (published or submitted) which are listed below in order of appearance with a short summary:

[1] Human-style interaction with a robot for cooperative learning of scene objects
S. Li, A. Haasch, B. Wrede, J. Fritsch, G. Sagerer

In research on human-robot interaction interest is shifting from uni-modal dialog systems to multi-modal interaction schemes. This is based on the observation that human-human interaction rarely relies on a single conceptual modality alone. Rather, humans make use of redundant and complementary cues provided by different information channels. Accordingly, robots offer the opportunity to model such embodied interaction in that they provide multi-modal sensory input and allow for multi-modal output. We present a system for human-style interaction with a robot that is integrated on our mobile robot BIRON. To model the dialog we adopt an extended grounding concept with a mechanism to handle multi-modal input and output where object references are resolved by the interaction with an object attention system (OAS). The OAS integrates multiple input from, e.g., the object and gesture recognition systems and provides the information for a common representation. This representation can be accessed by both modules and combines symbolic verbal attributes with sensor-based features. We argue that such a representation is necessary to achieve robust and efficient information processing.

[2] Developing a Contextualized Multi-modal Corpus for Human-Robot Interaction
A. Green, H. Hüttenrauch, E. A. Topp, K. Severinson Eklundh

The purpose of this paper is to describe the development process and challenges involved when collecting and annotating a corpus which is used in the research on cognitive robots in the European project Cogniron. One important aim is to support the development of natural language user interfaces. In the long run we aim to enable users to train the robot to perform a wide range of tasks that are not preprogrammed – using a multi-modal style of interaction. Our corpus contains task oriented conversation with a robot prototype, an ActivMedia PeopleBot, in a so-called home tour scenario developed in the project. In the scenario the user and robot move around in a home-like environment and the user names objects and locations using a combination of speech and gestures. We are striving to collect data from many different sources in order to be able to contextualize the modalities that are being used for interaction, i.e., communicative actions: speech and gesture and other actions related to the task. To our knowledge our corpus is unique because of its domain (task-oriented human-robot communication) and the multidimensionality of data that is collected and interlinked. So far we have used our corpus in the design process to evaluate the system from a usability perspective, to analyze miscommunication and to analyze users' positioning and task strategies.

[3] Integrating Miscommunication Analysis in Natural Language Interface Design for a Service Robot
A. Green, B. Wrede, S. Li, K. Severinson Eklundh

Natural language user interfaces for cognitive robots should provide an interaction that is perceived as smooth and intuitive to users. We present an analysis of miscommunication in 12

sessions (three hours) of human-robot dialog and dicuss the way design implications can be integrated in the development process. A large part of the miscommunication is related to users' (mis-)understanding of system's functions. This implies that the robot needs to provide relevant information about its states and capabilities, together with an efficient strategy for priming user behavior.

[4] An agent-based multi-modal dialog system for Human-Robot Interaction
S. Li, B. Wrede, G. Sagerer

Dialog systems for mobile robots operating in the real world should enable mixed-initiative dialog style and handle multi-modal information involved in the communication. Most dialog systems developed for mobile robots today, however, are often system-oriented and have limited capabilities. We present an agent-based dialog system that enables mixed-initiative, multi-modal dialog style. The first evaluation results fo this system indicate that these capabilities positively effect the interaction between human users and our robot as a whole.

[5] Do you like this robot? The role of robot behavior, robot personality and user personality
B. Wrede, S. Buschkämper, S. Li

We analyzed if users are able to assign personality traits to a robot and which factors influence liking of the robot. It turned out that users had no difficulties in judging the robot's personality. The analysis of the ratings revealed that the robot's dialog behavior (basic vs. verbose) had an effect on the personality ratings of the robot. Furthermore, the robot's behaviour, aspects of the robot's perceived personality and the user's personality contributed independent proportions of variance in explaining liking of the robot. These results indicate that the personality of users as well as of robots should be taken into account when designing human-robot interfaces.

Human-style interaction with a robot for cooperative learning of scene objects*

Shuyin Li, Axel Haasch, Britta Wrede, Jannik Fritsch, Gerhard Sagerer
Faculty of Technology
Bielefeld University
33594 Bielefeld, Germany
{shuyinli, ahaasch, bwrede, jannik, sagerer}@techfak.uni-bielefeld.de

ABSTRACT

In research on human-robot interaction the interest is currently shifting from uni-modal dialog systems to multi-modal interaction schemes. We present a system for human-style interaction with a robot that is integrated on our mobile robot BIRON. To model the dialog we adopt an extended grounding concept with a mechanism to handle multi-modal in- and output where object references are resolved by the interaction with an object attention system (OAS). The OAS integrates multiple input from, e.g., the object and gesture recognition systems and provides the information for a common representation. This representation can be accessed by both modules and combines symbolic verbal attributes with sensor-based features. We argue that such a representation is necessary to achieve a robust and efficient information processing.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Intefaces—*Natural language*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Tracking, Object recognition*; H.2.5 [**Heterogeneous Databases**]: Heterogeneous Databases

General Terms

Management

1. INTRODUCTION

Multi-modality is one of the most important features that characterize human-human social interaction. Based on this observation, designers of computer systems have been trying to integrate multi-modal input and output mechanisms in human-machine interactions to enable more intuitive operations for human users. Given the huge amount of existing multi-modal systems with quite different notions of multi-modality we first want to discuss two important aspects of modality intuitiveness to clarify our position. In general it is assumed that the degree of intuitiveness of a modality will determine the smoothness and efficiency of the interaction it facilitates. Thus, the question is what modalities are the most intuitive for users? We argue that the answer is of evolutionary nature and is application-dependent. For example, the mouse has become the major modality in human-computer interaction for many users. For these users the mouse is probably one of the most intuitive ways to operate a computer although it has little to do with the natural communication channels they use in human-human communication such as speech. In contrast, many other users prefer to write commands directly. We can imagine that future generations will find other modalities more intuitive than a mouse. The feeling of intuitiveness in computer operation is thus continuously changing. Even the same people may judge the intuitiveness of one modality differently when operating different computer systems. For example, people tend to anthropomorphize mobile robots when interacting with them. We argue that this is even more the case for application areas where a robot is supposed to assist and accompany humans in a social environment such as private households. Our envisioned robot is supposed to "live" in a private household. Our long term goal, therefore, is to endow it with social capabilities so that it can become some sort of "companion". This means that we do not envision it to serve humans in a master-slave manner as people tend to suppose. Rather, a robot companion should *cooperate* with humans to achieve certain goals. One basic function of such a Robot Companion is to be able to learn interactively about its environment. We therefore devised the *home-tour* scenario within our project where a human user is supposed to show his/her home to a newly purchased robot.

Based on our goal to design a robot that can be accepted by the user as a companion we argue that the interaction with such machines should be in a human style. We define the term *human style modalities* as *multi-modal communication channels that humans are biologically equipped for and (learn to) use from their birth.* Typical examples are speech and gestures. These modalities differ from other modalities like mouse and keyboard in that they are learned nat-

*This work has been partially supported by the European Union within the 'Cognitive Robot Companion' (COGNIRON) project (FP6-002020) and by the German Research Foundation within the Collaborative Research Center 'Situated Artificial Communicators' as well as the Graduate Program 'Task Oriented Communication'.

ICMI'05, October 4–6, 2005, Trento, Italy.

urally and without the use of artificial devices. In contrast, we define artificial modalities that are commonly used for human-computer interaction as *virtual modalities* because their effect is a virtual one which is only observable by humans via an artificial interface (e.g., the computer display). Thus, since people tend to anthropomorphize robots by expecting human-like abilities and attributes we conclude that robots should be endowed with human-style modalities for interaction.

A further aspect concerns the knowledge representation. Consider our home-tour scenario where the interaction will involve a high degree of deictic activity as the user will point to diverse objects in the environment. Thus, when the user points to his/her computer and explains "This is my computer", the robot should be able to recognize the user's gesture, find the computer and associate the symbolic name "computer" with a visual representation. This knowledge should be stored in a multi-modal way in order to be retrievable from different modules for further interactions. Psychological theories of knowledge representation in humans suggest that the symbolic name of an object is associated with its sensory features like its image or haptic characteristics (e.g. [3]). When activating the name of an object other features of the object are also activated. This indicates that the cognitive representation of objects in humans is multi-modal and therfore allows for multi-modal processing of information. In order for a robot to cooperate with a human, it should therefore be able to also process multimodal information to build a representation similar to that of its human communication partner to support a better mutual understanding. We therefore developed a multi-modal representation scheme.

To summarize our position shortly: The intuitiveness of modalities needed for operation of computers and machines has an evolutionary aspect and depends on the individual applications. In our Robot Companion domain we are interested in human-style multi-modality that should be considered for both, communication channels and the representation of knowledge. Its impact is of functional and technical nature.

In this paper we will first present the multi-modal processing strategies of the Dialog System (section 4) and the Object Attention System (section 5) followed by a detailed description of our multi-modal representation scheme in section 6. Results in the form of a dialog example will be given in section 7.

2. RELATED WORK

While there is an increasing interest in multi-modal interfaces there is only a very limited number of applications that use human-style modalities based on an integrated multimodal knowledge representation.

That multi-modal cues are beneficial for increasing the robustness in human-robot interaction has been shown for example in [13] where communication errors are detected by using not only speech recognition scores but also by including laser data to infer the presence or absence of communication partners and noise sources. Repair actions also involve the use of multiple modalities by either driving around to actively search for a communication partner or by offering buttons as alternative communication channels in noisy environments. More human-style interactions have been suggested in [1] by modeling a *naturalness support behavior*. This behavior includes verbal strategies by inserting filler phrases, as well as non-verbal reactions such as nodding or head turning as reactions to environmental noise. The authors report positive reactions by the users but extensive evaluations still remain to be done. The benefits of using multimodalities in a learning scenario have been demonstrated in several applications. For example, the robot Leonardo [2] can learn the names of buttons when a human communication partner points them out by verbal and deictic instructions. Leonardo also learns specific interactions with these buttons by demonstration. However, while the interactive capabilities of Leonardo are quite realistic the underlying representations are simple and no new objects can be learned. An impressive system running on a mobile robot is presented by Ghidary et al. [6]. The robot is able to learn objects by analyzing speech commands and hand postures of the user. The user gives verbal information about the object's size and can describe the spatial relations between objects, e.g., by phrases like 'left of my hand'. The rectangular views of the learned objects are stored in a map representing the robot's environment and can be used for later interactions. Although the interaction system is very limited and the resulting map is rather coarse, this system can be compared to our approach as it also builds up a long-term memory about objects in the environment. However, we focus on a more detailed representation of objects and their later recognition in order to support natural human-robot interaction going beyond simple navigational tasks.

In general there is a tendency to either focus on building a robust representation while neglecting interaction smoothness or vice versa. We argue that in order to build a robot that is able to be perceived as a companion it needs both, a more natural interaction based on an integrated multi-modal representation.

3. OVERALL SYSTEM

The scene acquisition system described in this paper is being implemented on our mobile robot BIRON. BIRON's hardware platform is a Pioneer PeopleBot from ActivMedia with a pan-tilt camera for face tracking and object and gesture recognition, stereo microphones for speaker localization and speech recognition, and a SICK laser range finder for locating legs of potential communication partners. The overall architecture [10] of BIRON is based on a hybrid control mechanism and has three layers: a reactive, an intermediate and a deliberative layer (see Fig. 1). Modules that are responsible for reactive feedback of the system are set on the reactive layer: the *Person Attention System* detects potential communication partners and the OAS detects objects that users refer to. Since these are purely data-driven processes they belong to the reactive layer. Modules responsible for higher-level processing that involve top-down, expectation-driven strategies such as a planner or the Dialog System, are located on the deliberative layer. The Scene Model, which contains a multi-modal representation of the objects that the system has observed and can be seen as an intermediator between the Dialog System and the OAS, is consequently located on the intermediate layer. The communication between modules is carried out via XCF (XML Enabled Communication Framework). The system is centrally controlled by the so-called *Execution Supervisor* on the intermediate layer. It coordinates the module operations and makes sure that neither the reactive layer mod-

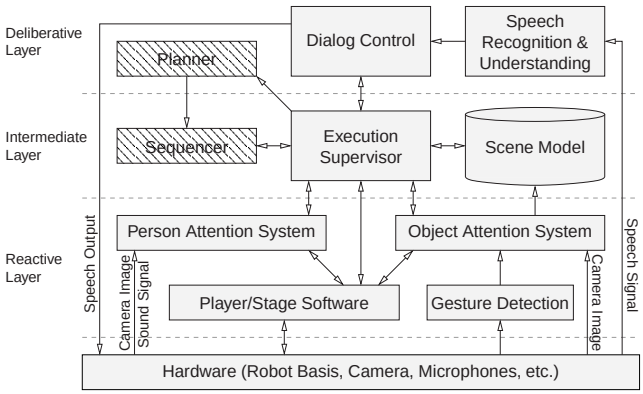

Figure 1: Overview of the BIRON architecture (optional modules currently not implemented are drawn in grey).

ules control the deliberative layer modules nor vice versa. Instead, it exerts control by taking into account the overall system state. This architecture allows for both fast reaction to dynamic environmental changes and extensive high-level planning and reasoning activities.

Since we focus on the interaction capabilities of BIRON, we do currently not integrate a planner and a sequencer. The Person Attention System establishes the basis for the interaction with users but is not involved in the process of resolving object references. In the following we therefore only describe the interfacing between the Dialog System, the OAS and the Scene Model.

4. THE DIALOG SYSTEM

The dialog system of BIRON is responsible for carrying out interactions with the user including handling miscommunications [16], guiding the discourse, and transferring user utterances to internal command for the robot control system to execute tasks. A dialog is made up of contributions from the dialog partners. Two central questions of dialog modeling are therefore (1) how to represent individual contributions represented and (2) how to represent the dynamic change of the dialog state represented which is triggered by individual contributions successively. In subsection 4.1 and 4.2 we are going to present our answers to these two questions. In section 4.3 we focus on the mechanism that the implemented dialog model provides for the integration of speech and visual input. A more detailed account on this integration will be given in section 5.

4.1 The structure of a contribution

Conversants contribute to a dialog in a multi-modal way. McNeill [12] investigated the relationship between speech and simultaneous conversational gesture and claims that the production of them are motivated by one single semantic source, the so-called "idea unit". Inspired by this finding, we represent the conversants' contribution as the so-called "interaction unit" that includes two important stages of the language production process. The structure of the interaction unit is illustrated in Fig. 2. An interaction unit has two layers: a *domain layer* and a *conversation layer*. The *domain layer* mirrors the cognitive activities of a dialog participant that motivate language production: If the interaction unit represents an utterance to be produced by the

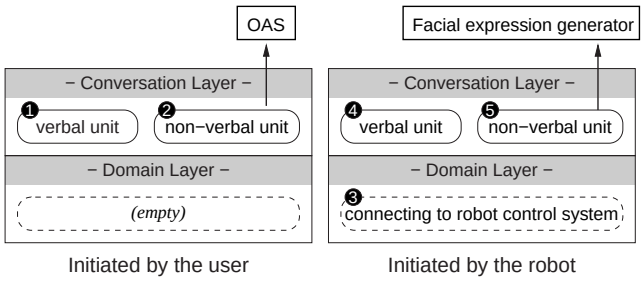

Figure 2: Processing flow in interaction units

robot itself the domain layer is where the Dialog System accesses the robot's control system or knowledge base. If the interaction unit represents an utterance of the user the domain layer remains empty because we do not make any assumptions about the user's cognitive activities behind the language front. In future work this may be replaced by a user model. The *conversation layer* transforms the intention that is created based on the results of these cognitive activities to language. For example, based on a successful follow behavior of the robot that is reported to the domain layer by the robot control system, the conversation layer formulates and synthesizes a message such as "OK, I follow you". The conversation layer consists of two units: a verbal and a non-verbal unit. Based on the intention that results from the domain layer they are responsible for generating output in verbal and non-verbal way, respectively.

The precondition of language production is successful language perception. Before reacting, i.e., before creating his/her own interaction unit to produce a contribution, a conversant first needs to understand the semantic meaning of his/her dialog partner's contribution by studying the verbal and non-verbal unit on the conversation layer of the dialog partner's interaction unit. Therefore the processing in the interaction unit should start from this language perception phase. Figure 2 illustrates how the robot processes user contributions. In our system, the user's verbal information as delivered by the Speech Understanding System initiates the creation of a user interaction unit (❶). In case that the user's intention can not be fully recognized by the verbal unit, the system will consult the visual perceptual module via OAS in the non-verbal unit (❷) and fuse these multi-modal information on the user conversation layer of the interaction unit. Once the user intention is fully recognized, the system creates an interaction unit for itself and tries to provide acceptance: the system first formulates and sends commands to the robot control system or the knowledge base on the domain layer (❸) and then generates verbal and non-verbal output on the conversation layer (❹, ❺) after receiving the execution results. Currently, we have implemented the visualization of facial expressions as the only non-verbal output. Thus, the integration of speech and visual information is mainly performed on the conversation layer of the interaction unit.

In the whole language perception and production process problems may occur, e.g., the semantic meaning of the user interaction unit cannot be resolved or the desired task cannot be executed by the robot system. These problems cannot be handled in a single interaction unit, new interaction units are necessary. Now the question arises as to how to organize the individual interaction units.

4.2 The grounding mechanism

The interaction units have to be organized in a dynamic way since every new contribution that is added to the dialog changes the dialog state. Our dialog model is inspired by the common ground theory of Clark [5]. According to this theory a dialog is carried out in the way that one participant presents an account (presentation) and the other issues the evidence of understanding of the account (acceptance). The grounding process is complete and both dialog participants can go on with a new account only if the acceptance is available. Dialog systems that implement this psychological model ([15], [4]) differ in their way of defining grounding units (the units of the discourse where the grounding takes place) and the organization of these units. We take exchanges in the style of adjacency pairs [14] as the grounding units. These exchanges consist of two interaction units that are initiated by the two dialog partners, respectively. The first interaction unit is the presentation and the second one is the acceptance, e.g., the first interaction unit represents a question and the second one the answer. To organize them we introduce four grounding relations between exchanges: (1) *default*: introducing a new task. A grounded default exchange has no further effect on the grounding of its preceding exchange. (2) *support*: clarifying in case of an ungroundable account. After a support exchange is grounded its initiator will try to ground the preceding one again that is updated with the new information. For example, clarification question in case of an incorrect speech recognition result. (3) *correct*: correct the previous account. As support, if such an exchange is grounded the initiator will try to ground the updated preceding one again. (4) *delete*: delete the previous account. If such an exchange is grounded, all the previous ungrounded exchanges can be deleted.

Each contribution is analyzed in terms of whether it is a presentation or an acceptance. If it is a presentation, then we also need to find out its grounding relation to the preceding exchanges. A presentation initiates the creation of an exchange that is put onto the top of a stack while an acceptance completes the top exchange of the stack. When the top exchange is complete it is popped. Additionally, actions like updating the preceding exchange can be triggered according to these relations. As long as there is an incomplete exchange on the top of the stack, the conversant other than the initiator of the exchange's presentation will try to ground it. The implemented dialog system enables us to handle clarification questions (as an exchange with support relation to its preceding exchange) and take initiative that is motivated by the robot control system. For example, in case of technical problems of the robot control system the implemented dialog system initiates an interaction unit to report this problem to the user. It does this by encoding the error message into its domain layer and generating output to the user on the conversation layer.

4.3 Resolving object references

In the following we detail how the Dialog System and the OAS cooperate to resolve object references in the user's utterances.

According to [9] there are three types of informational relations between gesture and speech: *reinforcement*, *disambiguation* and *adding information*. In our work, we focus on the "adding information" relation. When people use gestures to complete the meaning of their utterances they mostly indicate this intention in the utterance. For example, if a user says "This is my green mug" while pointing at a mug, the word "this" serves as a cue for the listener that he/she is using a gesture to specify the concrete location of the mug. But in case of the subsequent utterance "The mug is my favorite one" the listener usually does not expect a gesture but will search mentally in the dialog history which cup might be meant. According to these two different cases the Dialog System activates either the OAS or the Scene Model to resolve the object reference. This process can be illustrated as a UML activity diagram (see Fig. 3).

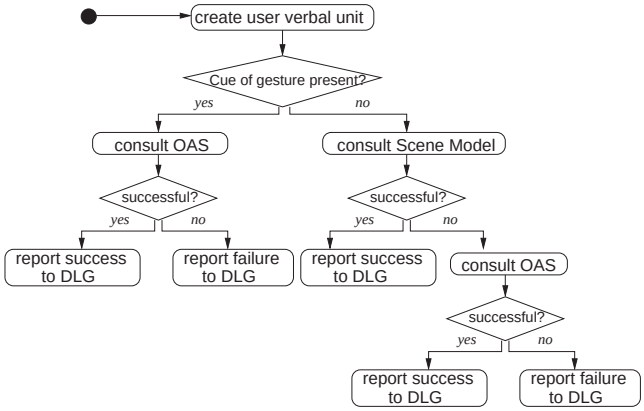

Figure 3: Resolving object references in user's verbal input (OAS: Object Attention System, DLG: Dialog System)

If there are any cues of the involvement of a gesture in the user's verbal input, e.g., the word "this" in the example above, this will be explicitely pointed out by the Speech Understanding System. The Dialog System interprets this hint as evidence that the user's verbal unit needs to consult the non-verbal unit. In the non-verbal unit the Dialog System activates the OAS by sending it the request to resolve the object reference "my green mug". The OAS activates the gesture recognition module. A successful gesture recognition result helps the OAS to orient the camera towards the position of the user's hand which enables the OAS to confine the Region Of Interest for its search of a green mug. This search is carried out, as the case may be, either by an appearance-based object recognizer or with the help of salient object features as described in section 5. Subsequently, the OAS sends the search result back to the Dialog System. In case of a positive result the OAS also updates the Scene Model with both symbolic and visual information about the new object "green mug". According to this result the Dialog System creates a system interaction unit (cf. Fig. 2) to provide acceptance for the user's input by either acknowledging or by initiating verbal repair.

If there is no evidence of the involvement of a gesture in the user's verbal input but only some objects to be identified (such as the "mug" in the example "The mug is my favorite one") the Dialog System will first try to find a corresponding entry in the Scene Model. The query is constructed with all the features of the object present in the current verbal input; in this case, the owner of the cup and his/her relation to this object (favorite). If the object can be found in the Scene Model the Dialog System finishes its processing on the conversation layer of the user interaction unit and creates an

acceptance to the user's input; if this object is not registered in the Scene Model, the Dialog System activates the OAS to find it in the current scene. This process is described in detail in the following section.

5. OBJECT ATTENTION SYSTEM

In order for a system to be able to acquire knowledge about objects in its environment it needs a mechanism to focus its attention on those objects that the user is currently talking about. In our context we define attention as the ability to select and concentrate on a specific stimulus out of all stimuli that are provided by the environment while suppressing others. The Object Attention System (OAS) therefore needs to coordinate the visual processing results (which currently consist of deictic gestures, object recognition results, and visual object features) and making them accessible for the Dialog System by storing them in the multi-modal Scene Model.

The OAS is activated when the user is verbally referring to an object and the Speech Understanding System has determined that either a gesture is expected or that the robot has to interact with an object autonomously. In order to acquire visual information about objects, BIRON uses an active camera with a maximal opening angle of view of about 50 degrees horizontal and 38 degrees vertical which facilitates only a limited field of view. It can therefore be necessary to re-orient the camera to relevant parts of the current scene which the user refers to. Once the robot has focused its attention on such a so-called *Region Of Interest*, the acquisition of information about this object, like position or view, can be completed. The Region Of Interest is that part of the image that has been specified by a gesture and contains the distinctive feature verbally specified by the user. Our assumption is that it contains an image of the object if the object is known to the robot. If it is unknown the Region Of Interest encloses the verbally specified visual feature (e.g. the "round thing" or a color).

The collected object data then has to be added to the robot's knowledge base which must allow retrieving stored object information and updating already stored data. Also, additional information given verbally by the user has to be stored. As this knowledge base, subsequently named as *Scene Model*, is crucial for the interaction between the Dialog System and the OAS because it represents BIRON's long-term memory, it is described in detail in section 6.

In addition to the maintenance of the Scene Model for memorizing tasks, the OAS needs to take care of the coordination of verbal information, gestures, and salient object features (e.g., color, shape, etc.) perceived by the camera, as well as the control of the hardware components like the camera during the object attention phases. This is realized by a *Finite State Machine* (see Fig. 5) where the input is mainly provided by the camera, the Dialog System (cf. section 4), the Speech Understanding System, and the gesture recognition component [7].

A crucial distinction that has to be made during the processing of multi-modal information is that between objects that are already known to the robot and those that are not (see Fig. 4). This is because in the latter case the OAS will have to establish (or 'learn') a first link between the verbal symbols describing the object and the percepts while in the former case the object needs to be retrieved from the database.

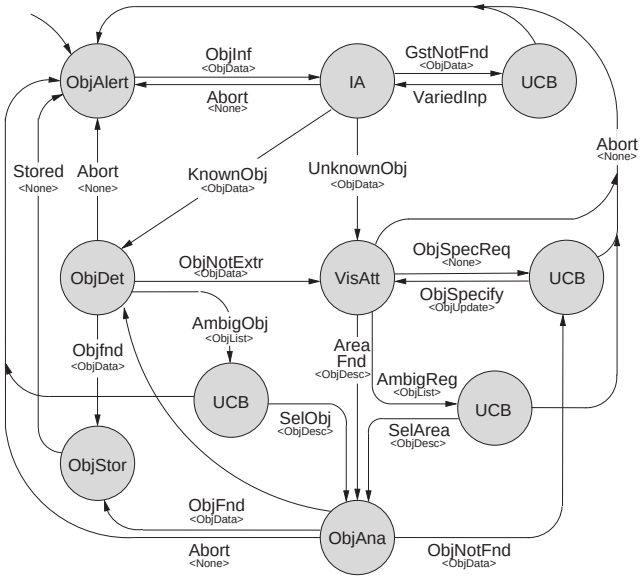

Figure 5: The Finite State Machine of the Object Attention System.

In both cases the OAS is activated on demand by the Dialog System if a gesture is expected or an access to the Scene Model by the Dialog System has failed afore. At this moment, the Finite State Machine (see Fig. 5) will be in the idle state *Object Alertness* (ObjAlert). Once the OAS is provided with data by the Dialog System, the Finite State Machine changes to the *Input Analysis* (IA) state. Now, the gesture recognition component is activated and provides the OAS with the user's hand coordinates and the direction of the corresponding pointing gesture. Thus, an area within the camera image is selected as the Region Of Interest. In case the Dialog System sends a description of the object (e.g., type, color, owner, etc.) to the OAS, a query to the Scene Model is initiated in order to check whether the object type is already known. In the following, we will describe in detail the processing for the case when the object type is known to the robot. The more complex process for the case of unknown objects will be exemplified subsequently.

5.1 Previously known objects

Suppose the user specifies an object type that the system has already stored in its Scene Model. In this case the Scene Model will return all object entries that match the symbolic description of the specified object. In order to verify if one of the returned objects is indeed the object the user refers to, the OAS will need to search for the object in the real scene and compare it with the stored image pattern. This search involves an object detection process for which we are currently using a simple appearance-based object recognizer that is only suitable for a very limited object scenario. It is based on the fast Normalized Cross-Correlation (NCC) algorithm described in [11] which is a simple but fast algorithm that is sufficient for our task at hand. However, in order for the system to work reliably in a more unstructured environment, as for example a real *home-tour* scenario a more sophisticated object recognizer will be needed.

After all appropriate image patterns have been retrieved for the known object type, the Finite State Machine switches

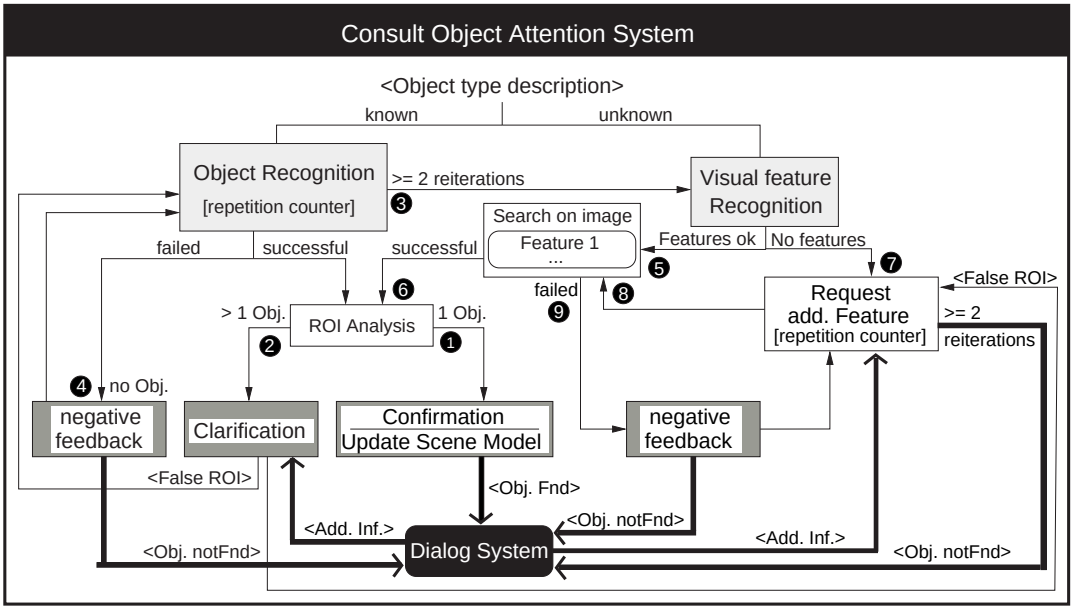

Figure 4: Schematic description of distinction between known and unknown objects

from the Input Analysis state to the *Object Detection* (ObjDet) state. Within the Object Detection state the OAS uses the retrieved image patterns (e.g., for cups) in order to feed them to the object recognizer. At the same time the camera is re-oriented based on the hand coordinates and the pointing direction that are provided by the gesture recognition component. Also, the position of the hand is used to determine the relevant Region Of Interest. Based on this information, the object detection process is initiated by scanning the Region Of Interest for patterns similar to those provided by the database. If an object is found by this procedure a confirmation message is sent to the Dialog System (cf. Fig. 4 (❶)) and the Finite State Machine switches to the *Object Store* (ObjStor) state. In this state the position of the object in the scene is updated in the Scene Model. Finally, the Finite State Machine returns to the Object Alertness state and the OAS awaits new orders.

If two or more objects of the same type are found in the real scene during the detection phase (cf. Fig. 4 (❷)) in the Object Detection state, the Finite State Machine will switch to the *User callback* (UCB) state. This means, that a message is sent to the Dialog System to clarify which of the found objects was meant by the user. After the Dialog System has provided a more detailed description, the Finite State Machine switches to the *Object Analysis* (ObjAna) state. In this state a new Region Of Interest is determined based on the information from the gesture detection and the extended verbal information. Now the Finite State Machine returns to the Object Detection state and initiates a new search. This cycle is performed until the object is found, or the user aborts the action within the User callback state but at most two times. Then, the Finite State Machine switches to the *Visual Attention* (VisAtt) state (cf. Fig. 4 (❸)). If no object is found in the Object Detection state (cf. Fig. 4 (❹)) this means that the user is referring to an unknown object which is supposedly similar to the description of objects previously retrieved from the Scene Model. However, since the object is unknown the Finite State Machine switches to the Visual Attention state, that is also used for the localization of unknown objects if two reiterations have been reached (cf. Fig. 4 (❸)).

5.2 Unknown objects

If no object detection is possible because no object entry matching the user's specification has been found in the Scene Model the OAS will search for salient features in the camera image such as colors or shapes by applying different filters that detect salient visual object features as specified by the user. We call these filters *attention maps* following the terminology of [8] where a similar technique is used. The use of these attention maps is coordinated within the Visual Attention state and can help to select Regions Of Interest. The appropriate attention map is selected based on the verbal information (e.g., the color) given by the user (cf. Fig. 4 (❺)).

Once a region matching the search criteria (i.e., color) is found within the Region of Interest by the attention map it is selected within a bounding box (cf. Fig. 4 (❻)). This bounding box is supposed to contain a view of the retrieved object (e.g., a blue cup) and is stored in the Scene Model (cf. Fig. 4 (❶)). Additionally, a confirmation message is sent to the Dialog System.

If the verbal information given by the user is insufficient to determine a Region Of Interest, that is if no visual descriptions are given that can be found by the attention map, (cf. Fig. 4 (❼)), the Finite State Machine changes to the User callback state. In the User callback state the OAS sends a request to the Dialog System in order to get more information (e.g., shape, position, ...) about the object which the user refers to. When the user has given a more specific description which is sent by the Dialog System to the OAS, the Finite State Machine returns to the Visual Attention state (cf. Fig. 4 (❽)). The User callback state is also reached if more than one Region Of Interest is found (cf. Fig. 4 (❷)). Then, the OAS asks the Dialog System to re-

solve this ambiguity. As soon as the OAS has determined the Region Of Interest, the Finite State Machine switches to the Object Analysis state to acquire the position of the object by means of the hand position of the user. Next, the Finite State Machine switches to the Object Store state and stores the extracted view and the position of the object in the Scene Model (cf. Fig. 4 (❶)). Then, the Finite State Machine returns to the Object Alertness state to await new orders.

If no Region of Interest is found during the search for visual object features, the Finite State Machine switches to the User Callback state and returns a negative response to the Dialog System. In parallel, the OAS asks the Dialog System for a more detailed object description and re-initiates a second search on the image (cf. Fig. 4 (❾)). If for a second time no Region Of Interest is found, the OAS sends a message to the Dialog System, that the search for the referenced object was not successful and returns into its idle state to await new orders from the Dialog System.

6. REPRESENTATION

Information acquired by the Dialog System and the OAS in the ongoing interaction with a user must be stored in an appropriate way. Because the same information from different modalities require different ways of representation the management of such a multi-modal database is a non-trivial task. Our approach to such a database, that we call *Scene Model*, is based on the concept of an active memory [17] since it uses intrinsic processes which allow not only a simple access to the data but also provides intelligent maintenance functionalities. One of the intrinsic processes for example enables the autonomous removal of obsolete information about objects (e.g., the position of a cup three months ago). This *forget* mechanism is quite essential for our application since the robot's environment is continously changing.

Within our work we have extended the functionality of the active memory in order to be able to handle the different modalities by storing the same data in different formats. This information that might seem to be redundant at first glance is necessary because the Scene Model is used as BIRON's long-term memory and both Dialog System and OAS access the data stored in it. For example, consider the color of an object: the Dialog System may store its value in form of a character string, e.g.,"blue"; but after finding it in the current scene the OAS may need to store its color value based on the *Hue Saturation Intensity* (HSI) color model. The same holds true for the position of an object, which the Dialog System would store symbolically as "on the table" while the OAS would store its coordinates. The coordinates describing the position of an object are obviously quite useless for the Dialog System when the user asks "where is my blue cup?". On the other hand, the OAS would not be able to handle the value "blue" when it has to find a blue cup in the camera image.

Consequently, the Scene Model needs to include a component that is able to convert the format of data, the so-called *Modality Converter*. The Modality Converter is a simple yet powerful mechanism that is not only able to convert single object features like the color. It can also search the data base and will return whole object entries matching given descriptions. This may be necessary for example when the OAS fails to detect an object by matching all memorized views against the current camera image and the object's color is not yet known to the OAS. Then, in order to extend the search for visual object features, the OAS sends a request for the object's color to the Dialog System. Subsequently, a search for the newly given color can be performed, after an appropriate conversion of the Dialog System's response is received by the OAS.

For the conversion process the Modality Converter uses a lookup table (cf. Table 1) that contains for every stored *predicate name* (e.g., color, relation, ...) two attribute fields, in particular a *symbolic description* as well as a *visual feature description*. Since the HSI values might vary for a distinctive verbally named color, the corresponding value field can contain specific values as well as ranges of values. Depending by which module a query is sent the Scene Model returns automatically the adequate description if available. For instance, if the query origins from the Dialog System the Converter will automatically return the attribute "Symbolic".

Predicate name	**Symbolic**	**Visual**
Color	red	330..20,0..1.0,0..1.0 0.0,1.0,1.0
Color	green	135,1.0,0.7
Relation	O_1 under O_2	$O_1.y < O_2.y$

Table 1: Lookup table of the Modality Converter

Note that this converter is a very powerful tool since it does not only convert pre-defined symbol-value pairs but it is also able to learn new associations. In sum, three different responses to queries are possible. ① The converter finds an entry where all data fields match the attributes of the specified object. In this case, it will return the value suitable for the inquiring component. ② The converter finds no valid entry because there is no correspondence of the entry in the other modality as for example for the symbolic name "transparent" which does not have an HSI equivalent. ③ The converter finds no valid entry because it is not yet complete. This usually occurs when after a search for visual object features a new HSI value is stored in the Scene Model for which no symbolic name is yet known. Then, the Dialog System will ask the user for the color name of the Region Of Interest.

7. RESULTS

In order to illustrate our results we present a dialog example where the resolution of object references is involved. In this example, the user asks the robot to pay attention to a mug. Fig. 6 illustrates the dialog flow, the operations of the modules underlying the robot output and the content of the Scene Model in the first, second and third column respectively. We assume that the robot has already recognized the user as Thomas.

In the utterance U1 the word "this" indicates a possible accompanying gesture which can help to specify its meaning. The Dialog System therefore sends a request to the OAS to search for the object mug (❶). Upon this request the OAS will first query the Scene Model for an object of the type mug in order to provide a template to the object recognizer (❷). In our example, no such object is stored in the Scene Model (❸). The OAS now switches to its second searching strategy: search with salient features of the object in the current scene. But since neither salient fea-

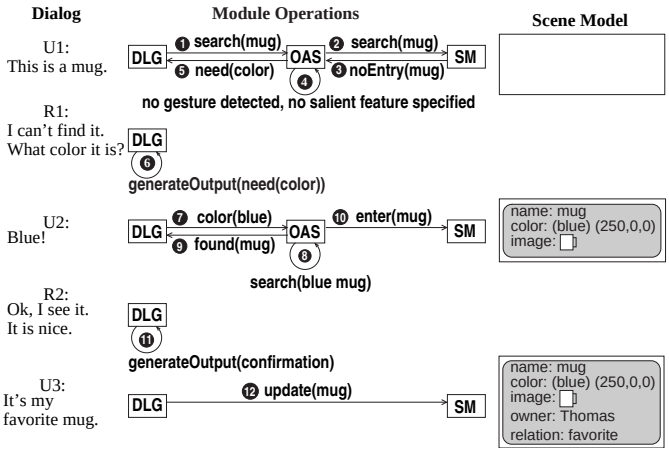

Figure 6: A dialog example (U: User; R: Robot; DLG: the Dialog System; OAS: the Object Attention System; SM: Scene Model)

tures of the mug were specified, nor a gesture was found in the current scene (❹), the OAS informs the Dialog System about its need of a salient feature, namely the color (❺). The Dialog System will then generate the output R1 to get the requested information from the user (❻). Thomas answers the question and the value of the color is sent from the Dialog System to the OAS (❼). With this information the OAS successfully finds the object mug (❽) and informs the Dialog System about this result (❿). At the same time the multi-modal information about the mug is entered in the Scene Model by the OAS (⓫). The Dialog System can now generate a confirmation (⓬) as feedback to the user. In U3 the user specifies two further features of the mug, the owner ("mine") and a description ("my favorite"). This information is directly entered into the Scene Model by the Dialog System since there is no evidence of the involvement of a user gesture, which would indicate that a new object is being specified, and there is only one entry in the Scene Model that matches the described object.

8. CONCLUSION

In order for a mobile robot to be able to communicate in a human-style it needs processing and representation strategies that can deal with multi-modal information. We therefore integrated a Dialog System using multi-modal interaction units with an Object Attention System that is able to resolve object references. The interaction between these modules is based on a multi-modal representation, the Scene Model, which stores the acquired scene information and provides a Modality Converter that not only converts information from one modality to another but also can learn associations between data from different modalities such as color names and HSI values. This powerful mechanism allows on the one hand to use a pre-defined knowledge base while it is on the other hand capable of adapting to new environments by learning new objects and salient features.

The current system still has some limitations. Firstly, the robot cannot yet learn new words, this means, the robot can only learn objects with known symbolic names. A solution of this problem can be adding a mechanism into the DLG that can store and reuse new words once they are spelled by the user. Secondly, a global 3D coordinate system representing the absolute position of an object in relation to the room is not yet integrated into the OAS. The consequence is, the robot can only find the object again if it is in the position as it learned the object. Thirdly, no navigation system is implemented for the robot so that the robot cannot autonomously move from one location to another to find an object. These limitations are also the motivation for our future work.

9. REFERENCES

[1] K. Aoyama and H. Shimomura. Real world speech interaction with a humanoid robot on a layered robot behavior control architecture. In *IEEE Int. Conf. Robotics & Automation*, pages 3825–3830, 2005.

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

[3] J. S. Brunder. *Towards a Theory of Instruction*. Norton, New York, 1966.

[4] J. E. Cahn and S. E. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[5] H. H. Clark, editor. *Arenas of Language Use*. University of Chicago Press, 1992.

[6] S. S. Ghidary, Y. Nakata, H. Saito, M. Hattori, and T. Takamori. Multi-modal interaction of human and home robot in the context of room map generation. *Autonomous Robots*, pages 169–184, 2002.

[7] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer. A multi-modal object attention system for a mobile robot. In *Proc. Int. Conf. on Intelligent Robots and Systems*, 2005.

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998.

[9] J. M. Iverson, O. Capirci, E. Longobardi, and M. C. Caselli. Gesturing in mother-child interactions. *Cognitive Develpment*, 14(1):57–75, 1999.

[10] M. Kleinehagenbrock, J. Fritsch, and G. Sagerer. Supporting advanced interaction capabilities on a mobile robot with a flexible control system. In *Proc. Int. Conf. on Intelligent Robots and Systems*, pages 3649–3655, 2004.

[11] J. P. Lewis. Fast template matching. In *Proc. Conf. on Vision Interface*, pages 120–123, 1995.

[12] D. McNeill. *Hand and Mind: What Gesture Reveal about Thought*. University of Chicago Press, 1992.

[13] P. Prodanov and A. Drygajlo. Decision networks for repair strategies in speech-based interaction with mobile tour-guide robots. In *IEEE Intern. Conf. Robotics & Automation*, pages 3052–3057, 2005.

[14] E. Schegloff and H. Sacks. Opening up closings. *Semiotica*, pages 289–327, 1973.

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

[16] D. Traum and P. Dillenbourg. Miscommunication in multi-modal collaboration. In *Proc. AAAI Workshop on Detecting, Repairing, And Preventing Human–Machine Miscommunication*, 1996.

[17] S. Wrede, M. Hanheide, C. Bauckhage, and G. Sagerer. An Active Memory as a Model for Information Fusion. In *Proc. Int. Conf. on Information Fusion*, pages 198–205, 2004.

Developing a Contextualized Multimodal Corpus for Human-Robot Interaction

Abstract for LREC-2006

Anders Green	Helge Hüttenrauch
Elin Anna Topp	Kerstin Severinson Eklundh

{green, hehu, topp, kse}@nada.kth.se

Interaction and Presentation Laboratory
KTH School of Computer Science and Communication
100 44 Stockholm, Sweden

Introduction

The purpose of this paper is to describe the development process and challenges involved when collecting and annotating a corpus which is used in the research on cognitive robots in the European project Cogniron[1]. One important aim is to support the development of natural language user interfaces. In the long run we aim to enable users to train the robot to perform a wide range of tasks that are not preprogrammed – using a multimodal style of interaction.

Our corpus contains task oriented conversation with a robot prototype, an ActivMedia PeopleBot, in a so-called Home-Tour scenario developed in the project (cf. Green, Hüttenrauch, & Severinson Eklundh, 2004). In the scenario the user and robot moves around in a home-like environment and the user names objects and locations using a combination of speech and gestures.

We are striving to collect data from many different sources in order to be able to contextualize the modalities that are being used for interaction, i.e., communicative actions: speech and gesture and other actions related to the task. To our knowledge our corpus is unique because of its domain (task-oriented human-robot communication) and the multidimensionality of data that is collected and interlinked.

There are initiatives to collect corpora for multimodal interfaces (e.g., Knudsen, Dykjær, & Bernsen, 2001; Schiel, Steininger, & Türk, 2002) but few that are targeted for robotics (e.g., Bugmann, Klein, Lauria, & Kyriacou, 2004; Wolf & Bugmann, 2005). Koide et al (2004) have collected and analyzed interaction statistics to investigate human reactions to specific robot behaviors (Koide et al., 2004). Other uses of corpus data include observations of user behavior, e.g., gaze behavior, to evaluate human engagement in interaction (Sidner, Kidd, Lee, & Lesh, 2004).

So far we used our corpus in the design process to evaluate the system from a usability perspective (Green et al., 2004), to analyze miscommunication and to analyse users' positioning and task strategies.

[1] www.cogniron.org

Data collection and annotation process

We recorded and transcribed 22 user sessions, each lasting approximately 15 minutes (5.5 hours of video). We used a Wizard-of-Oz set up that was perceived as realistic to the users. After the session we administrated a questionnaire. Audio from two different sources was collected: the sound from the wizard's video camera and the sound from the stereo microphones placed on top of the robot. This setup provided the overall picture of the robot and user acting together with the robot centric sound.

Several different types of data sources have been annotated and interlinked to support usability research and development of cognitive modules. The relation between dialogue acts and physical acts (cf. Traum, 2000) is especially interesting with respect to human-robot interaction together with the question of complexity of representation. We are aiming for a model where annotators need to make few decisions.

Speech annotation: The audio and video recordings are annotated up to what we could characterize as a baseline level: speech utterances and gestures have been transcribed and synchronized in order to provide a format that can be used to navigate the recordings. The synchronized transcriptions have been converted Anvil XML files (Kipp, 2004) allowing the sessions to be displayed in several layers.

Communicative acts: We are using a coding taxonomy to capture communicative acts that can be viewed as multimodal extension of the DAMSL coding schema (Allen & Core, 1997). Our extension of the schema currently involves deictic gestures, emblems, and iconic gestures. We are using a multi-layered style of annotation that allows for more detailed analyzes. Our approach is simular to Villaseñor et al (2000), who proposes the extension of DAMSL with the notion of contribution as participatory communicative acts, according to (Clark & Schaefer, 1989). In the paper we will describe and exemplify our schema further and its relation to current theories.

Positioning and spatial distance: We are annotating spatial formation , i.e., the dynamic aspects of spatial arrangements using a taxonomy based on Kendon's F-formation system (Kendon, 1990). This system is based upon the observation that certain patterns of posture and orientation between participants are maintained during interaction.

We are also coding interpersonal distances according to the classification proposed by Hall (1966). Social interaction is based upon and governed by four interpersonal distances: *intimate* (0–1.5 feet), *personal* (1.5–4 feet), *social* (4–12 feet), and *public* (>12 feet).

Both the F-formation system and social distances provide discrete representations for spatiality. Therefore we are also collecting and synchronizing laserdata and video recordings to be able to study this topic further.

Task: The schema used for task annotations is domain dependent and our intention at this point is that it should be used as background information for more focused analyses.

Scene overview: Images from four network web-cams that can be used to disambiguate the scene linked to the corpus using the timecode.

Laser range data: The data from the laser range finder is stored as raw data files with time stamps. This allows for development of different types of applications, e.g., tools for visualization or tracking algorithms.

Text descriptions and questionnaire data During the analysis of spatiality we also wrote down observations on events in the session, these text descriptions are time aligned so they can be easily retrieved. Questionnaires administered to users concerning their attitudes towards the system are also available.

Conclusions and future work

We have described the process of developing a contextualized corpus for human-robot interaction. By providing links to data sources, e.g., laser data and text descriptions and data that is annotated using well established taxonomies we aim to support activities related to the development of a cognitive robot. In the near future we will use this corpus in the development of adaptive models of users' style of communication and to study communicative behavior related to the spatial configuration of the robot and user.

References

Allen, J., & Core, M. (1997). *Draft of DAMSL: Dialog Act Markup in Several Layers.* webpage. (http://www.cs.rochester.edu/ research/cisd/ resources/ damsl/ RevisedManual/)

Bugmann, G., Klein, E., Lauria, S., & Kyriacou, T. (2004, March 10-13). Corpus-based robotics: A route instruction example. In *Proceedings of IAS-8* (pp. 96–103). Amsterdam, NL.

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

Green, A., Hüttenrauch, H., & Severinson Eklundh, K. (2004, 20-22 Sept). Applying the Wizard-of-Oz framework to Cooperative Service Discovery and Configuration. In *13th IEEE International Workshop on Robot and Human Interactive Communication RO-MAN 2004* (pp. 575–580).

Hall, E. T. (1966). *The Hidden Dimension: Man's Use of Space in Public and Private.* London, UK: The Bodley Head Ltd.

Kendon, A. (1990). *Conducting interaction - Patterns of behavior in focused encounters. Studies in interactional sociolinguistics*. Cambridge, NY, USA: Press syndicate of the University of Cambridge.

Kipp, M. (2004). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Knudsen, M. W., Dykjær, L., & Bernsen, N. O. (2001, 2 September). Surveys of multimodal data resources, annotation schemes and tools. In *Proceedings of the COCOSDA'2001 Workshop on Language Resources and Technology Evaluation - Technical, Global and Regional Perspectives* (pp. 135–146). Aalborg, Denmark.

Koide, Y., Kanda, T., Sumi, Y., Kogure, K., & Ishiguro, H. (2004, 28 Sept/ 2 Oct). In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004)* (Vol. 3, pp. 2500–2505).

Schiel, F., Steininger, S., & Türk, U. (2002, June). The SmartKom Multimodal Corpus at BAS. In M. G. Rodriguez & C. P. S. Araujo (Eds.), *Workshop on Multimodal Resources and Multimodal Systems Evaluation* (pp. 200–206). Las Palmas, Spain. (In association with the Third international conference on Language Resources and Evaluation LREC2002)

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to look: a study of human-robot engagement. In *IUI'04: Proceedings of the 9th international conference on Intelligent User Interfaces* (pp. 78–84). New York, NY, USA: ACM Press.

Traum, D. R. (2000). 20 Questions for Dialogue Act Taxonomies. *Journal of Semantics, 17*(1), 7–30.

Villaseñor, L., Mass, A., & Pineda, L. (2000, May). A multimodal dialog contribution coding scheme. In *The First EAGLES/ISLE Workshop on Meta-Description and Annotation Schemes for Multimodal/Multimedia Language Resources in conjunction with the Second International Conference on Language Resources and Evaluation LREC 2000.* Greece.

Wolf, J. C., & Bugmann, G. (2005, 12th-14th September). Multimodal Corpus Collection for the Design of User-Programmable Robots. In *TAROS 2005 Towards Autonomous Robotic Systems Incorporating the Autumn Biro-Net Symposium.*

Integrating Miscommunication Analysis in Natural Language Interface Design for a Service Robot

Anders Green
Kerstin Severinson Eklundh
Interaction and Presentation Laboratory
KTH Royal Institute of Technology
100 44 Stockholm, Sweden
{kse, green}@nada.kth.se

Britta Wrede
Shuyin Li
Faculty of Technology
Bielefeld University
33594 Bielefeld, Germany
{bwrede, shuyinli}@techfak.uni-bi.de

ABSTRACT
Natural language user interfaces for cognitive robots should attempt to reduce the occurence of miscommunication in order to be perceived as providing a smooth and intuitive interaction to its users. This paper will describe how we integrate miscommunication analysis in the design process. By analysing data from 12 sessions, where subjects interacted with a service robot in a home like environment, we arrived at a set of observations, e.g., that users misunderstand the robot's functionality; and that feedback sometimes is ill-timed with respect to the situation; we also observed that referencing objects is important with respect to lexical choice and deixis. The design implications from our analysis are that we need to equip our robots to provide more and relevant feedback with respect to the system's functionality. Another design implication is to explore strategies that prime the user to respond in a way that can be handled by the robot system.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: Natural language; I.2.9 [**Robotics**]: Operator interfaces

General Terms
Human Factors, Design

Keywords
Human-Robot Interaction, Miscommunication, Error handling, Dialouge design, Wizard-of-Oz

1. INTRODUCTION
The focus of the research presented here is to investigate models for how cognitive robots can work beside humans to assist them in their daily activities. A robot with cognitive capabilities needs an interface modality that ensures an intuitive and powerful way to reach the full potential of the system. It is generally believed that speech and gesture based interfaces provide a good model for human-robot interaction by offering an easy to learn, yet expressive way of communicating the user's goals and intention to the robot. Due to the situatedness and multimodal style of human-robot communication, miscommunication may occur along several dimensions, something which poses an even greater challenge than within the domain of speech interface research.

Figure 1: The user assumes an alternative mode of operation for gesture detection and holds up a magazine instead of placing it on a flat surface.

Therefore, one important goal when designing human-robot communicative systems is to provide interaction that is characterized by a low level of miscommunication, and is perceived as smooth and efficient by the user. Turning this into a research objective, our aim with this work is to gain a solid understanding of the causes of miscommunication in order to identify and handle it as it occurs during interaction. We are approaching this at the concrete level by evaluating a prototype dialog model that has been developed for the robot BIRON [15] using a Wizard-of-Oz type of setup [7] to collect data.

This paper is organized as follows. First we present related work and then we discuss how miscommunication analysis is used within our user oriented design process. Then we describe the setup of the study, the results from the miscommunication analysis and discuss the implications of it in terms of new dialogue design.

2. RELATED RESEARCH

Miscommunication can be defined as a state of misalignment between the mental states of agents involved in communication [17]. Either the speaker fails to produce the effect intended with the communicative acts issued or the hearer fails to perceive what the speaker intended to communicate. Analysis of miscommunication is sometimes referred to as "breakdown analysis". But a breakdown is only one extreme in a wide spectrum of possible miscommunication. It should be noted that we are not primarily interested in analyzing breakdowns *per se*, but symptoms of miscommunication that may lead to breakdowns.

There are few examples of focused miscommunication analysis in the field of human-robot interaction. Green et al [8] presented an explorative study of communicative errors relating them to the grounding model presented by Brennan & Hulteen [4]. Strategies for reducing miscommunication, i.e., using back-channel responses were discussed by Trafton et al [16]. In a recent study Breazeal et al [3] analyzed miscommunication in order to measure the effects of different non-verbal strategies that affect the efficiency and robustness of human-robot communication. Corpus collection e.g., [5] aimed at studying linguistic phenomena related to human-robot communication will typically contain data that can serve as a basis for miscommunication analysis.

The study of miscommunication has attracted interest within the spoken dialogue community. Here miscommunication is approached from different perspectives. Martinowsky & Traum [13] provide an example of how miscommunication analysis can be used to discuss human reaction to spoken dialogue systems. Symptoms of miscommunication are displayed at different levels in the exchange, e.g., as dialogue acts attempting to repair misunderstandings, as erroneous actions resulting from misunderstandings, and attitudinal responses to the exchange. They studied different phenomena that can be taken as indications of miscommunication, e.g., intonation, emphatic speech elliptic speech, vocatives, extra-linguistic signs and hyper-articulation.

There are also other more formal ways of classifying miscommunication, for instance Aberdeen & Ferro [1] who classified miscommunication using four features: the type of error; surface evidence available to the user (e.g., a repair act); the correction mechanism used (e.g., start over) and the outcome, whether the error was resolved or unresolved. Applied coherently this schema allows for using machine learning approaches to be used in the development process. While the data used by Aberdeen & Ferro [2] and Walker & Passonneau [18] was dialogue only, the multimodal character of human-robot communication complicates the discovery of error because users' gestures, posture and gaze behavior needs to be taken into account. Walker & Passonneau [18] were interested in more formal evaluation of dialogue systems providing the means of comparing different dialogue strategies. Their classification scheme was used to develop a dialogue parser and distinguishes between three orthogonally different levels of utterance classification: speech-acts, task-subtask dimension and conversational-domain dimension. The nature of development of human-robot communicative systems is that systems often are tightly connected with the domain and the particular robot platform and thus comparison between different systems is rarely a matter of concern.

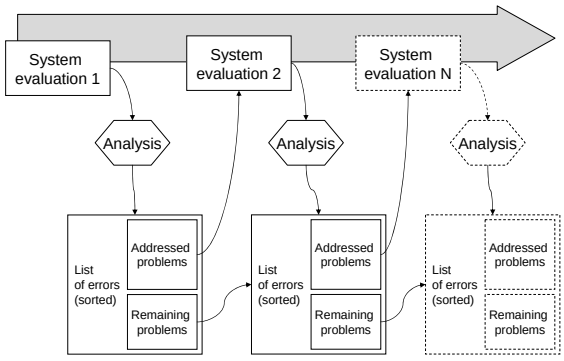

Figure 2: The role of miscommunication analysis in the design process.

3. MISCOMMUNICATION ANALYSIS

The way miscommunication analysis is being used within our design process can be illustrated with the schema depicted in Figure 2. When a prototype is evaluated it is analyzed from different perspectives. The result of an analysis focusing on miscommunication is a *list* of trouble spots. This list of identified problems can then be sorted according to different levels of priority. The severity of the problem needs to be weighed against the cost of addressing them. For instance, some problems can be addressed through changes in the dialogue design, e.g., by using a more effective prompting strategy or different wording, without the need to improve the backend components like speech recogntion and dialogue handling. Other problems require technical development, e.g., a more advanced microphone setup or new types of perceptual capabilities, for instance vision capability to handle pointing gestures.

Based on the sorted list of problems and the proposed solutions we can address problems in a systematic way. Thus, some problems can be addressed in the next version of the system but some will remain, either to the next level of system development or throughout the life time of the system. At the far end of this spectrum we find problems that require common sense knowledge or machine perception similar to human capability.

3.1 Purpose of the study

Methods for high-fidelity simulation, like the Wizard-of-Oz [7] framework provide an opportunity for different stakeholders in the development process to visualize and try-out the system without implementing it. In this framework a system that is being evaluated trough hi-fi simulation is fully or

partially simulated providing a situation where the user believes that she is interacting with a real system. This allows data collection in a realistic but yet controlled interaction situation.

In the context of the COGNIRON project we are interested both in improving the BIRON system and addressing more general topics of human-robot communication, such as aspects pertaining to the quality of communication.

3.2 Dialogue model

The prototype dialog model that has been adapted for the study described in the following sections is represented as a Finite State Machine (FSM) extended with a slot-filling mechanism [15]. This model has been implemented on the robot BIRON [15], an interactive robot system based on an ActiveMedia PeopleBot platform. A basic component of the robot system is the person attention system which enables the robot to focus its attention on one person. Based on this attention the robot can physically follow the person of interest and engage in verbal interactions. The heart of the system is the Execution Supervisor [15] which coordinates the communication between the different software components and represents the internal status of the system as an FSM.

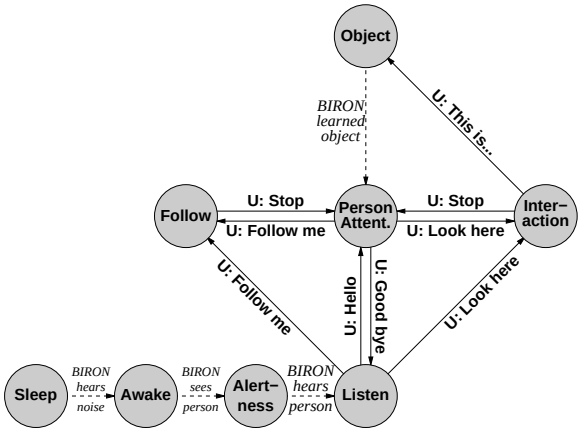

Figure 3: The dialog model described as a Finite State Machine.

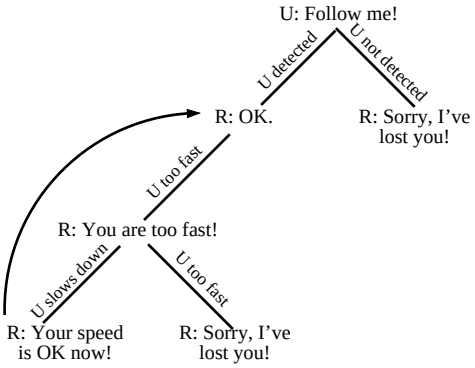

Figure 4: Sub-dialog of state "Follow".

The underlying FSM of the dialog system is identical to that of the central Execution Supervisor. The different states can thus be seen as the global "context" of the robot, indicating which task the robot is currently performing. Figure 3 illustrates the FSM of the dialog. Basically, the user can ask BIRON to do two things: either to follow her ("FOLLOW") or to pay attention to an object that is being shown to the robot ("INTERACTION"). This "interaction" state means, that the user has to "warn" the robot before she is showing an object. This is necessary because it needs to adjust its camera to the hands of the user to be able to detect a pointing gesture instead of focusing on the face. In each of these states, the robot can only do one thing, and the corresponding dialog between the user and the robot in this state will only focus on achieving this task.

These "sub-dialogs" are modeled by individual FSMs which also served as a basis for the specification of the robot behavior we investigated in this study (see Section 3.3.1). Figure 4 gives an example of the sub-dialog in the state "FOLLOW". Thus, if the user asks the robot to follow her, the robot will react depending on whether or not it detects the person. If the basic conditions for the task are met, that is if a person is detected, the robot will start following. However, once the robot notices that the distance to the user becomes too large it will try to correct this by informing the user. Similarly, the algorithm for the dialog exchanges in the interaction state are specified based on whether or not the system detects a gesture and an object.

3.3 Scenario and test procedure

We are envisioning a scenario where the user is teaching the robot important locations and objects using speech and gestures in combination, the so-called "Home Tour Scenario". Using the information given by the user the robot should then be able to perform tasks within the environment.

The scenario can be characterized as *Co-operative Service Discovery and Configuration*, stressing the way the user and robot are intended to engage in a joint effort to inform each other of relevant knowledge about the environment. This means that the user is able to configure the robot and discover what the it can do by actively providing information about *artifacts and regions* present in the environment (e.g., objects and locations) and; *trying the actions* that the robot can perform related to artifacts and regions (e.g., moving to places and finding objects).

For this study one requirement was that we would recruit users that were not familiar with robotics. We wanted to test the system in a setting that is as realistic as possible. To achieve this we decided to use the Wizard-of-Oz framework in a test area within the robot laboratory that is equipped with furniture normally found in a living room: a couch, a dining table with chairs, bookshelves, a TV set etc. Together with the "living-room" furniture, objects, like a fruit bowl, a remote control and some magazines were added to provide a set of objects that could be taught to the robot by users.

3.3.1 Adaptation of dialogue patterns

Thus we aim to test the system in a realistic but non-rigid manner, i.e., by providing a dialogue model with similar constraints as the implemented dialogue, but with enough robustness to allow for a habitable [10] dialogue system, i.e.

that there are no points in the system where the system lacks a model to handle input.

We used the dialogue patterns described in section 3.2 as a point of departure for the functions supported by the system:

Greeting:	responding to utterances like "hello robot".
Closing:	responding to utterances like "goodbye robot"
Person following	allowing the user to tell the robot to follow the user
Referencing locations and objects	responding to references to objects using speech (e.g. "this is an orange") together with deictic gestures

We are indeed interested in miscommunication but we do not want to provoke miscommunication. Thus we need to balance the system so that the aspects that want to test, i.e., particular dialogue design, are used in a way that makes them justice. Otherwise there is a risk that the user will experience an interaction filled with constant breakdowns due to causes that are unrelated to the dialogues system put up for evaluation.

3.3.2 Technical setup and test procedure

The robot system used in the data collection was an ActivMedia PeopleBot (similar to BIRON [15]). The robot was controlled by two researchers, also referred to as "wizards". The task of the wizards was divided in two roles: the navigator wizard and the dialogue wizard. The dialogue wizard provided the verbal means for the robot to reply to users commands using a speech synthesizer. The navigator wizard controlled the movements of the robot, including those of the camera, which is mounted on top of the PeopleBot.

Initially we performed a formative pilot study with a few staff members in order to fine tune the setup. In the next phase we recruited 22 test persons among students on the KTH campus. This means that there is a bias towards well-educated young people in the study, but since the aim of the study is primarily explorative we have accepted this circumstance. Upon arrival the subject was greeted by the test leader and offered a cup of coffee. Then the test leader informed the subject of the purpose of the study, without revealing that the wizards were controlling the system. Instead the wizards were described as "technicians" with the purpose of controlling the technical setup and making "online annotations". After the introduction the subject signed an agreement giving consent to storing of personal information. The users read the written instruction that explained the purpose of the study and the general functions the robot supported (see Section 3.3.1). The test leader also showed the follow behavior and pointed out an object to the robot. Then the robot was sent back to the standby position and the user could start the session. After about 15 minutes the session was ended on the initiative of the test leader. After the session we administrated a questionnaire assessing users' opinions of the interaction. Before leaving the subject was rewarded a cinema ticket voucher.

About 5.5 hours of video from the user sessions were recorded using a digital camcorder (MiniDV). Audio from two different sources was collected: the sound from the wizard's video camera and the sound from the stereo microphones placed on top of the robot. This setup provided the overall picture of the robot and user acting together with the robot centric sound.

4. RESULTS OF THE ANALYSIS

The video recordings from the first 12 user sessions have been transcribed and synchronized on the utterance level. We then printed out all dialogues and analyzed them by marking utterances that could be considered trouble spots, or symptoms of miscommunication. Then we checked the trouble spots in the video material to get a clearer picture of the characteristics of each situation. We then annotated the material using the Anvil [11] tool which provides visualization in the style of a musical score. We also generated a hypertext document that allowed us to move between a categorized and sorted list of trouble spots and the corpus texts to provide context.

In all we identified about 20 types of trouble spots, some occurred just once or twice but some were more frequent. We will limit our discussion to the categories that are frequent and that have had implications for the new design.

4.1 Users' system knowledge

At some points during the sessions exchanges that could be characterized as "trouble spots" either in terms of communication, i.e., where symptoms of miscommunication occurred or as in the exchange in the sequence $\mathbf{U}_1$-$\mathbf{U}_7$ (below), where a mismatch between the robot task capability and the tasks the user thinks the robot should handle.

$\mathbf{U}_1$ *stop robot* 31.306 - 31.935
$\mathbf{U}_2$ *turn around* 34.227 - 35.040
$\mathbf{R}_3$ *Stopped following* 35.778 - 36.895
$\mathbf{R}_4$ *Cannot do that* 38.154 - 38.904
$\mathbf{U}_5$ *Rotate* 40.579 - 41.317
$\mathbf{R}_6$ *Cannot do that* 43.118 - 43.835
$\mathbf{U}_7$ *follow me* 49.836 - 50.412

We have classified these errors as SYSTEM KNOWLEDGE referring to a trouble spot that can be attributed to what the user knows about the communicative capabilities of the system. This category also covers what may be considered requests for tasks that are out of the domain, e.g., praising the robot by saying "Good work robot". There are cases that are not clear cut, for instance when a user shows the robot and object by holding it in his hand instead of placing it on a flat surface. It is clear to the wizards that this is not an acceptable gesture, and it should also be clear to the user that holding objects should not work. The error can be said to belong in both categories, i.e., it is a communicative problem because the system fails to detect a gesture, but it is also a domain problem since the robot is supposed to handle objects on flat surfaces only.

In the sequence $\mathbf{U}_1$-$\mathbf{U}_7$ (above), several phenomena that can be characterized as symptoms of miscommunications occur. Initially the user is stacking commands: first the user is commanding the robot to stop, and then he asks the

robot to turn around. The response from the robot, i.e., that it has stopped following the user (utterance $\mathbf{R}_3$), in the contributions following the ones stacked by the user ($\mathbf{U}_1$, $\mathbf{U}_2$) is delayed about four seconds.

The stacking is in itself not a sign of miscommunication but the lack of feedback from the robot during the four seconds following the user's stop command can be regarded as an instance of the robot failing to make its contribution in a timely manner. It is worth noting that the robot actually stops right after the user has given the stop command, well before issuing the response "Stopped following" ($\mathbf{U}_3$). This renders the utterance spurious and ill-timed. On the other hand, when the robot utters "Cannot do that" ($\mathbf{R}_4$), referring to the users command "turn around" ($\mathbf{U}_2$), the user seems to interpret this as relevant to the exchange and attempts another adapted version of the turn command ($\mathbf{U}_4$).

After issuing the first turn command, "turn around" (in $\mathbf{U}_2$) the user has yet to discover that the robot cannot handle directive commands[1]. After the robot system has reported that it cannot perform the turn action, the user chooses to adapt his command by using the synonym "Rotate" ($\mathbf{U}_5$). After the robot has responded negatively to the second turn command (in $\mathbf{U}_6$) the user resorts to using what we believe the user considers a fallback command namely "follow me" ($\mathbf{U}_7$). The usage of the follow command in this type of situation is frequent in the material.

Another problem that is related to the users' erroneous inferences about the system's capability. Small objects, such as magazines, pens etc, were sometimes moved before being shown to the robot. Users tried to hold up objects in front of the camera. This was considered to be an error according to the task model and an a repair was issued by the communicator wizard. This behavior is illustrated in Figure 1. In the example below, the user is holding the pen while uttering $\mathbf{U}_{11}$. The repair $\mathbf{U}_{10}$ then influences the user's actions and a pointing gesture is issued.

$\mathbf{U}_8$ *this is a table* 119.256 - 120.323
$\mathbf{R}_9$ *Found dinner table* 136.353 - 137.520
$\mathbf{U}_{10}$ *this is a pen* 140.054 - 140.938
$\mathbf{R}_{11}$ *Rearrange the objects please* 146.797 - 147.897
$\mathbf{U}_{12}$ *this is a pen* 150.366 - 151.291
$\mathbf{R}_{13}$ *Found one object* 155.376 - 156.701

4.2 Feedback problems

We have noted several types of problems related to feedback in our data. Providing relevant and timely feedback essential to maintaining an orderly and well managed dialogue. We have identified problems related to timing, i.e., feedback is *ill-timed*, something which may render it incoherent, like in utterances $\mathbf{U}_{14}$-$\mathbf{U}_{16}$ (below). When the user utters "stop" ($\mathbf{U}_{14}$) and then tries to specifiy an object ($\mathbf{R}_{15}$) he is interrupted by the robot saying "stopped following" ($\mathbf{U}_{16}$). Issuing "stopped following" ($\mathbf{U}_{16}$) is thus non-relevant since the robot already stopped. At this point in dialogue this does not cause a breakdown but if the error occurs again, the user needs to adapt to the system's behavior, something that might affect the attitude towards the system.

[1]Except for "back" which we added to make it possible to get the robot free when it got stuck or to close.

$\mathbf{U}_{14}$ *stop* 122.355 - 122.804
$\mathbf{U}_{15}$ *this is a table* 125.819 - 127.233
$\mathbf{R}_{16}$ *Stopped following* 127.233 - 127.847

Another problem that occurred in the material was lack of feedback, i.e., the robot does not respond to the user's contribution before the user decides to make another contribution (cf. $\mathbf{U}_{17}$-$\mathbf{U}_{19}$) utterances. Here the time between repetitions may be used to provide a measure of the time limits for what the user will accept as a response time in the system [8].

$\mathbf{U}_{17}$ *this is a computer* 62.436 - 64.253
$\mathbf{U}_{18}$ *this is a computer* 68.746 - 70.380
$\mathbf{R}_{19}$ *Found one object* 76.857 - 78.257

4.3 Irrelevant response

In the following example the robot speech output overlaps ($\mathbf{U}_{21}$-$\mathbf{R}_{22}$) with the users command. This seems to lead to hesitation of the user in the later utterance ($\mathbf{U}_{24}$). This is a symptom of miscommunication but as a single instant it seems that it is not sufficient to cause a breakdown.

$\mathbf{U}_{20}$ *Stop* 122.355 - 122.804
$\mathbf{U}_{21}$ *this is a table* 125.819 - 127.233
$\mathbf{R}_{22}$ *Stopped following* 127.233 - 127.847
$\mathbf{R}_{23}$ *Found coffee table* 137.918 - 139.268
$\mathbf{U}_{24}$ *ahm* 141.081 - 141.738
$\mathbf{U}_{25}$ *please follow me* 144.195 - 145.254

4.4 Reference

In the manner the system was simulated we allowed for a "robust" object recognition system, meaning that the system would recognize any object given that it was small enough and placed on a flat surface. We had not added lexical entries in the simulated dialogue system meaning that we sometimes needed to reference objects with the word "object". This caused some concerns for some users since they, as we have interpreted their behaviour, were not sure about the status of the recognized object. In the following example the user attempts to get the system to name the object by saying "Identify object" ($\mathbf{U}_{28}$). Once again it is hard to classify this trouble spot. This exchange could also be related to the category System Knowledge. If we take on a strict logical perspective, the positive response of the user to the utterance "Is this the object" cannot justify that the robot actually found the newspaper. The information can be said to have been negotiated, but since there is no pointing capability apart from the general direction indicated by the front robot and the onboard camera, there is no way of indicating precisely which object has been detected.

$\mathbf{U}_{26}$ *this is a newspaper* 1013.560 - 1014.829
$\mathbf{R}_{27}$ *Found one object* 1020.069 - 1021.379
$\mathbf{U}_{28}$ *identify object* 1023.030 - 1024.488
$\mathbf{R}_{29}$ *Is this the object* 1029.984 - 1031.220
$\mathbf{U}_{30}$ *Yes* 1032.057 - 1032.610
$\mathbf{R}_{31}$ *Ok* 1035.993 - 1036.602

5. DESIGN IMPLICATIONS

The examples given in the previous section describe the most frequent types of communication difficulties in an embedded

human-robot-situation. Based on this data we identified four main aspects that need to be optimised in our dialog model: increase the information given to the user, prime the user to only use words known to the system, monitor the communicational success, and develop recovery strategies.

5.1 Information given to the user

Many problems arise because of the users' limited or erroneous knowledge of the system's functionalities and because the feedback given by the robot is not sufficient. This is related to the observations to the categories System knowledge (Section 4.1) and Feedback (Section 4.2).

Such problems are especially frequent in embodied conversations since the understanding of an utterance is heavily dependent on the sensory information which are the base of the robot's world model. This world model of the real world can thus be highly error-prone and therefore needs to be communicated to the user.

One obvious strategy is to provide information upon explicit questions by the user (e.g. "what can you do?" or "what now?"). However, this requires a thorough design of the answer, based on information optimization criteria (e.g. Gricean Maxims [9]) and initiative modeling.

More promising is therefore a more implicit strategy of giving more specific information when they are required, for example when a user command cannot be executed as in example $\mathbf{U}_{10}$-$\mathbf{U}_{11}$ ("rearrange objects") where the user *holds* up an object instead of pointing to it. Here, the user does not know how to solve the problem and gives up the task. In such cases more information that helps to solve the problem is necessary.

However, as the system itself does not have enough information to know exactly what the problem is – the system's problem is simply that it did not detect a gesture – the help needs to be based on prior knowledge about users' errors such as user studies. In this case the user has to be informed that the objects need to be on a flat surface. To be able to issue context dependent help in this manner, the system needs to know when it *misdetects* a gesture.

A further strategy is the use of additional non-verbal (mainly visual) feedback to provide faster or simultaneous information, i.e., without blocking the audio channel. For example, in the sequence $\mathbf{U}_1$–$\mathbf{U}_7$ the (redundant) feedback "stopped following" is given too late ($\mathbf{U}_3$) and completely unnecessarily since the robot has already stopped, bringing the interaction out of synchronization. In such cases, the execution of the task is a sufficient feedback signal. In our new dialog model each interaction unit is composed by a verbal and a non-verbal contribution and provides thus a convenient framework for using non-verbal feedback. However, non-verbal feedback is not appropriate for all tasks. Non-verbal reactions to instructions such as "This is a book" generally need much more time than the verbal reaction since the movement of the camera towards the target position requires a lot of computation time in order to detect the gesture and compute the goal position of the camera. Thus, based on time-measurements from real system interactions it is possible to group the robot's non-verbal reactions with respect to whether or not they are fast enough to replace the verbal feedback.

A second line of problems that can be tackled by giving non-verbal information relates to resolving references. In example $\mathbf{U}_{28}$ ("identify object") the user initiates a clarification dialog to make sure that the robot focuses on the object referenced by he user. Given the complexity of the task to resolve references to objects in the real world we ended up defining a very narrow menu-like clarification structure with the help of a visual feedback screen replacing a pointing device [12]. Reverting to a more restricted dialog structure in difficult communication situations is a well known strategy in dialog design [19]. Even if this strategy alleviates some problems related to providing feedback, increasing the conversational capabilities to make the interaction more natural remains a research challenge.

5.2 User priming

Speech recognition errors and errors related to language understanding were the easiest ones to detect and to react to by the wizards in the simulated system – needing only a feedback asking for repetition. However, we observed that they caused severe problems with respect to the communicational smoothness of the interaction.

In general, repeated speech recognition errors lead to a break-up of the current task by the user initiating a new task (e.g., utterance $\mathbf{U}_{25}$). These difficulties are much harder to detect than problems related to non-executable tasks since they are more implicit and can only be detected as a pattern ranging over several consecutive utterances. Thus, since detection and repair may pose severe challenges, a better strategy might be to avoid these problems at all. One main problem causing these difficulties lies in the use of out-of-vocabulary (OOV) words. This is because the user is not aware of the robot's lexical capabilities. However, theories about alignment (e.g. Pickering & Garrod [14]) in human-human communication predict that the speaking styles of communication partners will converge during a communication, affecting lexical choice as well as syntactic, prosodic and pragmatic structures. Thus, by only using words that are part of the passive lexicon of the speech processing system we can prime the user to use words known to the system rather than OOV words. This also applies to the syntactic structure of the utterances that the speech recognition system accepts. Based on this observation we follow the strategy of implicit priming as described by Yankelovich [19]. For example, upon the computer's self explanation "I can follow you" the user is much more likely to use the command "follow me" instead of "come here" or "move".

5.3 Monitoring communicational success

If an abrupt topic switch can not be averted it will still be important for the system to monitor the quality of the current interaction in order to adapt the strategies of the system, taking measures to increase the communicative success. We may for instance adopt a more restricted dialog structure, or we may provide detailed feedback as suggested above.

In future work, we will introduce a measurement of the communicational success that monitors the ongoing interaction

and detects patterns indicating troubles. In the new dialog model, we have defined a first basic version of such a measure by counting the number of system initiated repair utterances. A more sophisticated approach would be to collect as many potential features as possible, e.g. duration between utterances, emotional cues, expectation violations, topic progression etc. and compute a communication success rate by using pattern recognition methods or defining thresholds.

5.4 Recovery strategies

Even though our goal is to minimize communication difficulties there will always be trouble spots. For such cases it is important to provide recovery strategies that help to re-establish the communication if a breakdown occurs. In the Woz studies we observed that users develop their own strategies. One strategy that we observed several times in the material was the use of a "fallback" strategy, i.e., communicative actions that the users have learned is working robustly. The most prominent example of this is the use of the "follow me" command (e.g., utterance $\mathbf{U}_{25}$).

Another type of recovery that is necessary comes from the many attempts at using directive commands such as "turn around" etc. This has to do with the users' knowledge about the system (see Section 4.1).

In the further design process of the whole system it is therefore necessary to provide the system with small but robust functionalities such as directive commands (e.g., "back", "rotate left" etc), or offering sub-dialogues related to the current situation (e.g., User questions "what else can I do?" or "what do you suggest now?").

5.5 Discussion

Considering all these implications for our new dialog model, instead of the (rather system-oriented) FSM model, we will employ a model based on theories of grounding (e.g. [6]). This will allow us to react to the current situation in a more flexible way instead of using pre-defined situation patterns and responses.

This allows us to interpret the ongoing interaction with respect to grounding aspects and to design the feedback with respect to how well the communication proceeds. Thus, if the system can not accept a fact presented by the user, e.g. because of execution problems, the system will initiate a clarification dialog. Additionally, in the new dialogue model each contribution has a verbal and a non-verbal part allowing sharing of complementary information between modalities. Furthermore, it also allows to define different response strategies based on situational variables such as the communicational success or other available information.

6. CONCLUSIONS

We have described a model for how miscommunication analysis can be integrated in the design of the user interface for a robot with cognitive skills. We collected and analyzed dialogue data using a Wizard-of-Oz setup, simulating the movements and the dialogue system. The transcribed data was annotated with respect to miscommunication. We found miscommunication of different types and on different levels in the communication.

The most prominent type of miscommunication was related to the users' understanding of the capability of the system. The gulf between what the system can handle and what the users believe the robot is capable of, can be viewed from different perspectives.

First of all, users are not accustomed to cognitive robots at all. This means that the user is involved in a learning experience from the start. Miscommunication, according to Martinowsky and Traum [13], gives the users information about the boundaries of the system's capabilities, allowing the user to test hypotheses about the system allowing for learning to take place. For instance, when the users assume that the system can handle several similar types of directive commands because the function "backwards" was allowed. Another case where learning takes place, but where it is not a clear cut case is when the user is supposing some task capability that the robot cannot handle, for instance, when the user holds an object in his hand instead of placing it on a flat surface. Here communication works – the robot provides negative feedback or directions to the user – but the task cannot be performed.

Thus the design implication, that users need more and relevant information, related to specific situations can be seen as a way of increasing the opportunity for users to learn from instances of miscommunication. However, information given to the user needs to be relevant. By carefully modeling feedback provided by the system, e.g., based on communicative principles, like Gricean Maxims [9], we can provide information to the user but avoid an excessively talkative robot.

Miscommunication related to speech recognition and natural language understanding affects the smoothness of communication. This leads to the design implication that we should attempt to prime the user into selecting lexical terms and syntactic structures that the system can handle. Priming can be considered a well established practice in more classical approaches to designing speech based system. This is one example how practices from human-computer interaction can be used in designing human-robot communication. However, establishing what types of strategies, e.g., as discussed by Yankelovic [19] that can be transferred easily and if some strategies will be invented remains a topic of research. One such area regards how multimodal feedback can be used in the robot interface to reference objects, for instance as proposed in Section 5.1 using visual feedback devices to disambiguate object references.

Adapting to communicative strategies of different users is an area where we miscommunication analysis serves an important purpose. The collected data can be used in various ways to train or inform models for measuring communicative success, something that well motivates the thorough annotation procedure.

If we think about miscommunication analysis as a step performed as an integral part of the design process, the way we have ordered the list of errors has influenced what design implications that we found worth concentrating our efforts and resources on when developing the next version of the system. We cannot hope to catch all errors in one system iteration, but should aim to get rid of the most severe prob-

lems every time we revise the system. They key here is to prioritize the list of identified problems so that we address them in an order that will have the most positive impact on the amount of problems that recur in later versions of the system.

We should see miscommunication analysis, and the resulting list of trouble spots, both as a way of increasing the understanding of the particular system being evaluated and as a way of tracking recurring and difficult problems. With this perspective on miscommunication analysis we are both providing a basis for improved design in the short term as well as providing challenging problems for research on human-robot communication.

7. ACKNOWLEDGMENTS

The work described in this paper was conducted within the EU Integrated Project COGNIRON ('The Cognitive Robot Companion' – `www.cogniron.org`) and was funded by the European Commission Division FP6–IST Future and Emerging Technologies under Contract FP6-002020.

8. REFERENCES

[1] J. Aberdeen, C. Doran, L. Damianos, S. Bayer, and L. Hirschman. Fnding errors automatically in semantically tagged dialogues. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 124–128, 2001.

[2] J. Aberdeen and L. Ferro. Dialogue patterns and misunderstandings. In *Proceedings of Error Handling in Spoken Dialogue Systems*, pages 17–21, Château dÓex, Vaud, Switzerland, August 28-31 2003.

[3] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Edmonton,Alberta, Canada, 2005.

[4] S. E. Brennan and E. Hulteen. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems*, 8:143 – 151, 1995.

[5] G. Bugmann, S. Lauria, T. Kyriacou, E. Klein, J. Bos, and K. Coventry. Using verbal instruction for route learning. In *Proceedings of 3rd British Conference on Autonomous Mobile Robots ande Autonomous Systems: Towards Intelligent Mobile Robots (TIMR'2001)*, Manchester, April 2001.

[6] H. H. Clark and S. E. Brennan. Grounding in communication. In L. R. Teasley, J. Levine, and S.D., editors, *Perspectives on socially shared cognition*, pages 127 – 149, Washington, DC, 1991. Reprinted in R. M. Baecker (Ed.), Groupware and computer-supported cooperative work: Assisting human-human collaboration. San Mateo, CA: Morgan Kaufman.

[7] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of Oz studies - why and how. *Knowledge-Based Systems*, 6(4):258–256, 1993.

[8] A. Green and K. Severinson Eklundh. Task-oriented Dialogue for CERO: a User-centered Approach. In *Proceedings of 10th IEEE International Workshop on Robot and Human Interactive Communication*, Bordeaux/Paris, September 2001.

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

[10] K. Hone and C. Baber. Designing habitable dialogues for speech-based interaction with computers. *International Journal of Human Computer Studies*, 54(4):637–662, 2001.

[11] M. Kipp. *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Dissertation.com, Boca Raton, Florida, 2004.

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

[13] B. Martinovski and D. Traum. Breakdown in human-machine interaction: the error is the clue. In *proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16, 2003.

[14] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–225, 2004.

[15] I. Toptsis, S. Li, B. Wrede, and G. A. Fink. A multi-modal dialog system for a mobile robot. In *ICSLP*, volume 1, pages 273–276, Jeju, Korea, 2004.

[16] J. G. Trafton, A. C. Schultz, N. L. Cassimatis, L. M. Hiatt, D. Perzanowski, D. P. Brock, M. D. Bugajska, and W. Adams. Cognition and multi-agent interactions from cognitive modeling to social simulation. chapter Communicating and collaborating with robotic agents. Cambridge University Press, 2006.

[17] D. Traum and P. Dillenbourg. Miscommunication in multi-modal collaboration. In *In working notes of the AAAI Workshop on Detecting, Repairing, And Preventing Human–Machine Miscommunication*, pages 37–46, August 1996.

[18] M. A. Walker and R. Passonneau. Date: A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *In Human Language Technology Conference*, San Diego, March 2001.

[19] N. Yankelovich. How do users know what to say? *ACM Interactions*, 3(6), 1996.

An Agent-Based, Multi-modal Dialog System for Human Robot Interaction

Shuyin Li, Britta Wrede, Gerhard Sagerer
Faculty of Technology, Bielefeld University
Bielefeld Germany
shuyinli, bwrede, sagerer@techfak.uni-bielefeld.de

Abstract

Dialog systems for mobile robots operating in the real world should enable mixed-initiative dialog style and handle multi-modal information involved in the communication. Most dialog systems developed for mobile robots today, however, are often system-oriented and have limited capabilities. We present an agent-based dialog system that enables mixed-initiative, multi-modal dialog style. The first evaluation results of this system indicate that these capabilities positively effect the interaction between human users and our robot as a whole.

1 Introduction

Natural language is the most intuitive way to communicate for human beings (Allen et al., 2001). It is, therefore, very important to enable dialog capability for personal service robots which should help people in their everyday life. However, the interaction with a robot as a mobile, autonomous device is different than with many other computer controlled devices which effects the dialog system design. Here we want to first clarify the most essential requirements for dialog systems for human-robot interaction (HRI) and then outline state-of-the-art dialog modeling approaches to position ourselves.

The first requirement results from the *situatedness* (Brooks, 1986) of HRI. A mobile robot is situated "here and now" and cohabits the same physical world as the user. Environmental changes can have massive influence on the task execution. For example, a robot should fetch a cup from the kitchen but the door is locked. Under this circumstance the dialog system *must* support mixed-initiative dialog style to receive user commands on the one side and to report on the percepted environmental changes on the other side. Otherwise the robot had to break up the task execution and there is no way for the user to find out the reason.

Another challenge for HRI dialog management is the *embodiment* (Duffy and Joue, 2000) of a robot which changes the way of interaction. Empirical studies show that the visual access to the interlocutor's body affects the conversation in the way that non-verbal behaviors are used as communicative signals (Nakano et al., 2003). For example, to refer to a cup that is visible both to the speaker and the listener, the speaker tends to say "this cup" while pointing to it. The same strategy is considerably ineffective during a phone call. This example shows, multi-modal communicative cues must be taken into account for a HRI dialog system.

To enable mixed-initiative, multi-modal dialog style is also the ambition of many HCI dialog modeling approaches. McTear (2002) classified these approaches into three main types: *finite-state-based*, *frame-based*, and *agent-based*. The finite-state-based approach follows a simple script of prompts and users have to answer predefined questions relevant to the task. In the frame-based approach the fixed dialog context is represented as a set of parameters that need to be filled before the task can be initiated. Both approaches can only handle well-structured tasks and enable either system- or user-led dialog styles. In the agent-based approach, the communication is viewed as a *collaboration between two intelligent agents*. In such approaches, dialog planning is often integrated into the task planning process. For example, within the TRAINS/TRIPS project several complex dialog systems for collaborative problem solving have been developed (Allen et al., 2001). Here the dialog system is viewed as a conversatonal agent that performs communicative acts. During a conversation, the dialog system selects the communicative goal that is based on its current belief about the domain and the general conversational obligations. In such sys-

tems both the system and the user can ask questions, request clarification, etc. They thus enable mixed-initiative dialog style and can handle more complex tasks. However, it is difficult to port such systems to a new domain since the domain-specific task planning is a part of the system. To eliminate this disadvantage Allen (2001) proposed an *abstract problem-solving model*. In the HRI field, due to the complexity of the overall systems, usually finite-state-based and frame-based strategies are employed for dialog systems e.g., (Aoyama and Shimomura, 2005; Bischoff and Graefe, 2002) or only certain aspects of the communication are studied in-depth e.g., (Fry et al., 1998; Kanda et al., 2002). As to the issue of multi-modality, most dialog systems view it as an architectural issue: The nonverbal behaviors are implemented as extra module that runs more or less "independently" of the spoken dialog system, i.e., when necessary, the nonverbal processing is activated to help to resolve ambiguity (Rothkrantz et al., 2004) or to smooth the spoken dialog flow (Aoyama and Shimomura, 2005).

In this paper we present an agent-based dialog model for HRI. As described in section 2, the two main contributions of this model are the new modeling approach of Clark's grounding mechanism and the integration of multi-modality handling ability. In section 3 we outline the capabilities of the implemented system and in section 4 we present the results of the first system evaluation.

2 Dialog Model

We view a dialog as a collaboration between two agents. Agents are subject to common conversational rules and participate in a conversation by issuing multi-modal contributions (e.g., by saying something or showing some facial expression). In subsection 2.1 we show how we handle conversational tasks by modeling the conversational rules based on grounding and in subsection 2.2 we present how we model individual contributions to tackle the issue of multi-modality. In subsection 2.3 we put these two things together to complete the model description.

2.1 Grounding

One of the most influential theories on the collaborative nature of dialog is the common ground theory of Clark (1992). In his opinion, agents need to coordinate their mental states based on their mutual understanding about the current tasks, intentions, and goals during a conversation. Clark termed this process as *grounding* and proposed a contribution model. In this model, "contributions" from conversational agents are considered to be the basic component of a conversation. Each contribution has two phases: a *Presentation* phase and an *Acceptance* phase. In the Presentation phase the speaker presents an utterance to the listener, in the Acceptance phase the listener issues an evidence of understanding to the speaker. The speaker can only be sure that the utterance she presented previously has become a part of their common ground if this evidence is available.

Although this well established theory provides comprehensive insight into human conversation two issues in this theory remain critical when being used to model dialog. The first one is the recursivity of Acceptance. Clark claimed, since everything said by a conversational agent needs to be understood by her interlocutor, each Acceptance should also play the role of Presentation which needs to be accepted, too. The contributions are thus to be organized as a graph. However, this implies that the grounding process might never really end (Traum, 1994). The second critical issue is taking contribution as the most basic *grounding unit*. In Clark's view, the basic grounding unit, i.e., the unit of conversation at which grounding takes place, is the contribution. To provide Acceptance for a contribution agents may need to issue clarification questions or repair. But when modeling a dialog, especially a task-oriented dialog, it is hard to map one single contribution from one agent to a domain task since tasks are always cooperately done by the two agents (Cahn and Brennan, 1999). Traum (1994) addressed the first issue by introducing a finite-state based grounding mechanism and Cahn and Brennan (1999) used "exchanges"' as the basic grounding unit to tackle the second critical issue. We combine the advantages of their work and present a grounding mechanism based on an augmented push-down automaton as described below.

Basic grounding unit: As Cahn and Brennan we take *exchange* as the most basic grounding unit. An exchange is a pair of contributions initiated by the two conversational agents. They represent the idea of *adjacency pairs* (Schegloff and Sacks, 1973). The

first contribution of the exchange is the Presentation and the second contribution is the Acceptance, e.g., if one asks a question and the other answers it, then the question is the Presentation and the answer is the Acceptance. In our model, a contribution only represents *one* speech act. For example, if an agent says "Hello, my name is Tom, what is your name?" this utterances is segmented into three Presentations (a greeting, a statement, and a question) although they occur in one turn. These three Presentations initiate three exchanges and each of them needs to be accepted by the interlocutor.

Changing status of grounding units: Also as proposed by Cahn and Brennan, an exchange has two states: *not (yet) grounded* and *grounded*. An exchange is grounded if the Acceptance of the Presentation is available. Note, the Acceptance can be an implicit one, e.g, in form of "continued attention" in Clark's term. Taking the example above, the other agent would reply "Hello, my name is Jane." without explicitely commenting Tom's name, yet the three exchanges that Tom initiated were all accepted.

Organization of grounding units: As Traum we do not think that the Presentation of one exchange should play the role of the Acceptance of its previous exchange. Instead, we organize exchanges in a stack. The stack represents the whole ungrounded discourse: ungrounded exchanges are pushed onto it and the grounded ones are popped out of it. One major question of this representation is: *What has the grounding status of one exchange to do with the grounding status of the whole stack?* Jane's Acceptance of Tom's greeting has no apparent relation to the remaining two still ungrounded exchanges initiated by Tom. But in the *center embedding* example in Fig. 1, the Acceptance of B1 (utterance A2) contributes to the Acceptance of A1 (utterance B2). These examples show that the grounding status of the whole discourse depends on (1) the grounding status of the individual exchanges and (2) the relationship between these exchanges, the *grounding relation*. These relations are introduced by the Presentation of each exchange because they start an exchange. We identified 4 types of grounding relations: *Default*, *Support*, *Correct*, and *Delete*. In the following we look at these relations in more detail and refer to exchanges with relation *x* to its *immediately preceding exchange* (IPE) as "*x* exchange", e.g., Support exchange:

Default: The current Presentation introduces a new account that is independent of the previous exchange in terms of grounding, e.g., what Tom said to Jane constructs three Presentations that initiate three default exchanges. Such exchanges can be grounded independently of each other.

Support: If an agent can not provide Acceptance for the given Presentation she will initiate a new exchange to support the grounding process of the ungrounded exchange. A typical example of such an exchange is a clarification question like "I beg your pardon?". If a Support exchange is grounded its initiator will try to ground the IPE again with the newly collected information through the supporting exchange.

Correct: Some exchanges are created to correct the content of the IPE, e.g., in case that the listener misunderstood the speaker and the speaker corrects it. Similar to Support, after such an exchange is grounded its IPE is updated with new information and has to be grounded again.

Delete: Agents can give up their effort to build a common ground with her interlocutor, e.g., by saying "Forget it.". If the interlocutor agrees, such exchanges have the effect that all the ungrounded exchanges from the initial Default exchange till now are no longer relevant and the agents do not need to ground them any more.

Having looked at the effects of the grounding relations we can describe the grounding mechanism with an augmented push-down automaton (APDA). This automaton (see Fig. 2) is augmented in so far that transitions can trigger actions and variable number of exchanges can be popped or pushed in one step. There are five states in this APDA and they represent the fact what kind of ungrounded exchange is on the top of the stack. Along the arrows that connect different states the input (denoted as I[...]), the resulting stack operation (denoted as S[...]) and the possible action that is triggered (denoted as A[...]) are given. The input of this automaton includes Presentation with one of the four relations to its IPE (e.g., "defaultP" stands for "Default Presentation") and Acceptance.

As long as there is an ungrounded exchange at the top of the stack, the addressee will try to ground it by providing Acceptance, unless she deletes its valid-

A1: What do you think about Mr. Watton?
B1: Mr. Watton? our music teacher?
A2: Yes. (accept B1)
B2: Well, he is OK. (accept A1)

Figure 1: An example of center embedding

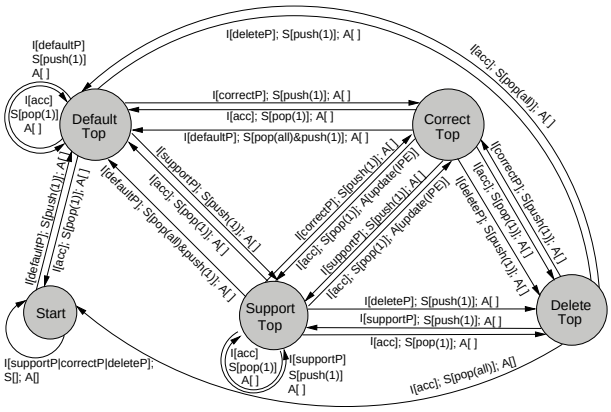

Figure 2: Grounding APDA

ity. For the reason of space, we only can explain the APDA with the center embedding example in Fig. 1. Contribution A1 introduces a question into the discourse which initiates a Default exchange, say Ex1. This exchange is pushed onto the stack. Instead of providing Acceptance to A1, contribution B1 (Mr. Watton? Our music teacher?) initiates a new exchange, say Ex2, with grounding relation Support to Ex1 and is pushed onto the stack. Then contribution A2 (Yes.) acknowledges B1 so that Ex2 is grounded and popped out of the stack. The top element of the stack is now the ungrounded Ex1. Since Ex2 supported Ex1, the Ex1 is updated with the information contained in Ex2 (The music teacher was meant) and B2 then successfully ground this updated Ex1.

In our model, every exchange can be individually grounded and contributes to the grounding of the whole ungrounded discourse by acting on the IPE according to their grounding relations. This way we can organize the discourse in a sequence without losing the local grounding flexibility. This model allows both the user and the system as equal conversational agents to easily take initiative or issue clarification questions. To implement this model, however, two points are crucial. The first one is the recognition of the user's contribution type: for every user contribution, the dialog system needs to decide whether it is a Presentation or an Acceptance. If it is a Presentation, the system needs further to decide whether it initiates a new account, corrects or supports the current one, or deletes it. This issue of intention recognition is a classical challenge for dialog systems. We present our solution in section 3. The second point is that the dialog system needs to know when to create an exchange of certain grounding relation by generating an appropriate Presentation and when to create an Acceptance. For that we need to first look at the structure of individual contributions more closely in the next subsection.

2.2 The structure of agents' contributions

To represent the structure of the individual contributions we take into account the whole language generation process which enables us to come up with a powerful solution as described below.

The layers of a contribution: What we can observe in a conversation are only exchanges of agents' contributions in verbal or non-verbal form. But in fact the contributions are the end-product of a complex cognitive process: language production. Levelt (1989) identified three phases of language production: *conceptualization*, *formulation*, and *articulation*. The production of an utterance starts from the conception of a *communicative intention* and the semantic organization in the conceptualization phase before the utterance can be formulated and articulated in the next two phases. Intentions can arise from the previous discourse or from other motivations such as needs for help or information. This finding motivates us to set up a two-layered structure of contributions. One layer is the so-called *intention layer* where communication intentions are conceived. For a robot the communication intentions come from the analysis of the previous discourse or from the robot control system. The other layer is the *conversation layer*. The communication intentions are formulated and articulated here[1]. These two layers represent the intention conception and the language generation process, respectively. We term this two-layered structure of contribution *interaction unit* (IU).

The issue of multi-modality: Face-to-face conversations are multi-modal. Speech and body lan-

[1] Since most robot systems use speech synthesizer for acoustic output which replaces the articulation process, only formulation is performed on this layer.

guage (e.g., gesture) can happen simultaneously . McNeill (1992) stated that gesture and speech arise from the same semantic source, the so-called "idea unit" and are co-expressive. Since semantic representation is created out of communicative intentions (Levelt, 1989) we assume the communication intentions are the modality independent base that governs the multi-modal language production. We, therefore, extend our structure above by introducing two generators on the conversation layer: one *verbal* and one *non-verbal* generator that represent the verbal and non-verbal language generation mechanism based on the communication intentions created on the intention layer. The relationship between these two generators is variable. For example, Iverson et al. (1999) identified three types of relationship between speech and gesture: *reinforcement*, *disambiguation*, and *adding-information*. In our work we focus on the adding-information relation, i.e., the gesture (generated by the non-verbal generator) contributes to the information conveyed in the speech (generated by the verbal generator). The structure of an IU is illustrated in Fig. 3.

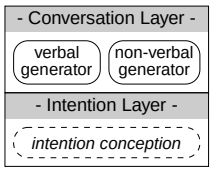

Figure 3: IU

Operation flow within an interaction unit: During a conversation an agent either initiates an account or reply to the interlocutor's account. The communication intentions can thus be *self-motivated* or *other-motivated*. For a robot, self-motivated intentions can be triggered by the robot control system, e.g., observed environmental changes. In this case, an IU is created with its intention layer importing the message from the robot control system and exporting an intention. This intention is transfered to the conversation layer which then formulates a verbal message with the verbal generator and/or constructs a body language expression with the non-verbal generator. Other-motivated intentions can be triggered by the needs of the on-going conversation, e.g., the need to answer a question, or be triggered by certain task execution results provided by the robot control system. The operation flow is similar to that of the self-motivation apart from the fact that, in case of intentions motivated by conversational needs, the intention layer of the IU does not import any robot control system message but creates an intention directly. Note, the IUs that are initiated by the robot and by the user have identical structure. But in case of user initiated IUs we do not make any assumption of their underlying intention building process and the intention layer of their IUs are thus always empty.

With the IUs, we can integrate the non-verbal behavior systematically into the communication process and model multi-modal language production (even if in the implemented system the image processing that is needed for the non-verbal behavior analysis is usually done by another robot component). The close coupling of intention and language production helps to achieve behavior consistency (Badler and Allbeck, 2000). We can also extend the conversation layer with a modality selection component that selects the best suitable modality for certain intentions to simulate more natural interaction behavior for a robot.

2.3 Putting things together

Till now we have discussed our concept of using a grounding mechanism to organize contributions and of representing individual contributions as IU. Now it is time to look at the still open point at the end of the section 2.1: when to create an IU as Presentation and when an IU as Acceptance.

Self-motivated intentions usually trigger the creation of an IU as Presentation with Default relation to its IPE. For example, if the robot needs to report something to the user it can create a Default exchange by generating an IU as its Presentation. The user is then expected to signal her Acceptance. Other-motivated intentions can, according to the context, result in either Presentation or Acceptance. To make the correct decision we developed criteria based on the *joint intention theory* of Levesque et al. (1990) which predicts that during a collaboration the partners are committed to a joint goal that they will always try to conform till they reach the goal or give up. Note, this does not mean that one will always agree with her interlocutor, but they will behave in the way that they think is the best to achieve the goal. This theory can be applied to human-robot dialog in a twofold sense: Firstly, a dialog can be generally seen as a collaboration as Clark proposed. Secondly, the human-robot dialog is mostly task-oriented, i.e., the human and the robot

work towards the same goal. With this theory in mind we describe how we process other-motivated contributions in the following.

The precondition of language production based on other-motivated intentions is language perception. Before reacting, i.e., before creating her own IU, an agent first needs to understand the intention conveyed by her interlocutor's IU by studying its conversation layer. We assume that agents first study the generated verbal information, if the intention can not be fully recognized here, one will further study the information provided by the non-verbal generator (e.g., a gesture) and fuse the verbal and non-verbal information. If the intention recognition is still unsuccessful, the agent can not provide Acceptance for the given IU. If she is still committed to the dialog she will issue a clarification question, i.e., she generates an IU as Presentation that initiates a Support exchange to the current ungrounded exchange. If the intention of her interlocutor is successfully recognized the language perception process ends and the agent tries to create her own IU. As described in subsection 2.2 the creation of the IU starts from the creation of an intention on the intention layer. In case of a robot, the dialog system accesses the robot control system and awaits its reaction to the conveyed information (e.g., a user instruction). Usually, a robot is designated to do something for the user, i.e., the robot is committed to the goal proposed by the user, so we define *the robot can only provide acceptance if the task is successfully executed*. In this case, the robot completes the current IU with the filled intention layer by generating an confirmation on its conversation layer. Afterwards, this grounded exchange can be popped from the stack. If the robot can not execute the task for some reasons, then the current exchange can not be grounded and the robot will take the current IU with the filled intention layer as another Presentation that initiates a Support or Correct exchange to the current ungrounded exchange, similar as the case in Fig. 1. The conversation layer of this IU can thus formulate something like "Sorry, I can't do that because..." and present a sorrowful face. This new Support or Correct exchange is pushed onto the stack. Figure 4 illustrates this process as a UML activity diagram.

In our model we only do general conversational planning instead of domain specific task planning.

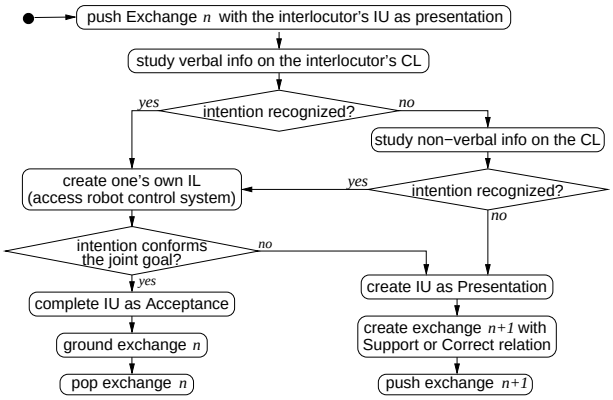

Figure 4: Handling other-motivated contribution (CL: Conversation layer; IL: Intention Layer)

What the dialog system needs to know from the robot control system is what types of intentions it can produce in terms of whether they start a new account, support or correct one, or delete it. Based on this information IUs are generated that operate according to the grounding mechanism as described in section 2.1.

3 Implementation

This dialog model was implemented for our robot BIRON, a personal robot with learning abilities. It can detect and follow persons, focus on objects (according to detected deictic gestures) and store collected multi-modal information into a memory. Our implementation scenario is the so-called *home tour*: a user shows a new robot her home to prepare it for future tasks. The robot should be able to "learn" features of objects that it was shown, e.g., its name, color, images etc. and to remember them (by putting this information into the memory).

The dialog system is linked to a speech understanding system which transforms parts of speech from a speech recognizer into semantic representation in a speech-act-based form. To recognize the user's intention, the dialog system classifies this input into 10 categories of three groups according to heuristics, e.g., *instruction, description*, and *query* initiate new tasks and thus a new Default exchange; *deletion* and *correction* initiate Delete and Correct exchanges that are related to early exchanges; and *confirmation*, *negation*, etc. can only be responses and are, therefore, considered as user's Acceptance

of exchanges that the robot initiates. The main part of the dialog system is the Dialog Manager that does all the grounding analysis and stack operations. The Robot Control Manager receives messages from the robot control system and calls the Dialog Manager to do relevant operations. This dialog system enables multi-modal, mixed-initiative dialog style and can handle complex clarification questions and repair. With the example dialog in Table 1 (U: user, R: robot), which is transcripted from a real interaction session between a user and BIRON, we describe the two most important features of our system.

Table 1: Dialog example with the extrovert BIRON

R1: Hello human! *(looking friendly)*
U1: Hello robot! Um, who are you?
R2: My name is BIRON. If you show me something I can remember it. *(looking friendly)*
U2: OK, um, look, this is a mug.
R3: I beg your pardon? *(looking puzzled)*
U3: This is a mug, umm, here.
R4: Sorry, can you use some other words? *(looking sad)*
U4: This is a mug!
R5: Sorry, I can't find it. Are you pointing to it with your finger? *(looking sad)*
U5: Yes!
R6: I know it is sometimes difficult with me, but don't feel discouraged! (*looking ashamed*)
U7: *(laugh)*
R8: OK, I've found it. It is really nice! *(looking happily)*

Taking Initiative and robot personality: Initiatives that a dialog system can take often depends on its back-end application. Since BIRON does not have a task planner which would be ideal to demonstrate this ability we implemented an *extrovert* personality for BIRON (additionally to its *basic* personality) that takes communication-related initiatives. The basic BIRON behaves in a rather passive way and only says something if user asks it. In contrast, the extrovert BIRON greets persons actively (R1 in Table 1) and remarks on its own performance (R6). When the robot control system detects a person in its vicinity the dialog system initiates a Default exchange to greet her. BIRON can also measure its own performance by counting the number of Support exchanges it has initiated for the current topic. Since the Support exchanges are only created if BIRON can not provide Acceptance to the user's Presentation (because it does not understand the user or it can not execute a task), the amount of the Support exchanges thus has direct correlation to its overall performance. On the other hand, the more Default exchanges there are, the better is the performance because the agents can proceed to another topic only if the current one is grounded (or deleted). Based on this performance indication BIRON does remarks to motivate users.

Resolving multi-modal object references: It happens quite frequently in the home tour scenario that the user points to some objects and says "This is a *y*". BIRON needs to associate its symbolic name (and eventually other features) mentioned by the user with the image of the object. The resolution of such multi-modal object references (U4-R8 in Table 1) is solved as following: the Dialog Manager creates an IU for the user-initiated utterance (e.g., "this is a cup") and studies the verbal and non-verbal generator on its conversation layer. In the verbal generator, what the pronoun "this" refers to is unclear, but it indicates that the user might be using a gesture. Therefore, the Dialog Manager further studies the non-verbal generator. The responsible robot control module is activated here to search for a gesture and to identify the object cup. If the cup is found in the scene, this module assigns an ID to the image and stores it in the memory. After the Dialog Manager receives this ID, the processing of the conversation layer of the user IU ends, the Dialog Manager proceeds to create its own IU to react to the user's IU. Problems with the object identification indicate failure of the intention recognition process on the user conversation layer. In this case, the Dialog Manager creates a Support exchange to ask the user which object she refers to and retries it if she dose not oppose (R8). This process is described in (Li et al., 2005) in detail.

4 Evaluation

The first evaluation of our dialog system was done in the context of the overall robot system. In summary, each of the 14 subjects who were not familiar with BIRON interacted with it for two consecutive runs and each run took approx. 5 minutes. Half the subjects interacted with the extrovert (Group-E), the other half with the basic (Group-B) version of the dialog. With this between-subject scenario we wanted to find out if and how the implemented communication-related initiative taking functional-

ity affects the users' perception of the interaction quality.

In total 28 runs the dialog system generated 903 exchanges with an average processing time of 11 ms. The analysis of the filled questionnaires showed more interesting results: 86% in Group-E as compared to 28% in Group-B stated that they 'liked' BIRON. Furthermore, only 28% of Group-E criticriticized BIRON was not giving enough feedback as compared to 57% in Group-B. In general, subjects of Group-E tended to describe the capabilities of X in a more positive way in open questions and to tolerate performance problems more easily than subjects in Group-B. Thus, although the robot performance in both conditions remained the same the extrovert dialog type did seem to positively affects the users' perception of the interaction quality. Of course evaluations of larger scale need to be carried out to investigate this effect in more detail, but this result is a strong evidence for the effectivity of the mixed-initiative dialog style that our system enables.

5 Conclusion

In this paper we presented an agent-based dialog system for HRI. The implemented system enables multi-modal, mixed-initiative dialog style and is relatively domain independent. The evaluation results provide strong evidence for the positive effect of these features. We will port this system to another robot system to enable the robot to take task-related initiative and to test it with more complex dialog.

References

J. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent. 2001. Towards conversational human-computer interaction. *AI Magazine*, 22(4).

K. Aoyama and H. Shimomura. 2005. Real world speech interaction with a humanoid robot on a layered robot behavior control architecture. In *Proc. Int. Conf. on Robotics and Automation*.

I. N. Badler and J. M. Allbeck. 2000. Towards behavioral consistency in animated agents. In *Proc. Deformable Avatars*.

R. Bischoff and V. Graefe. 2002. Dependable multimodal communication and interaction with robotic assistants. In *Proc. Int. Workshop on Robot-Human Interactive Communication (ROMAN)*.

R. A. Brooks. 1986. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23.

J. E. Cahn and S. E. Brennan. 1999. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*.

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

B. R. Duffy and G. Joue. 2000. Intelligent robots: The question of embodiment. In *BRAIN-MACHINE*, Ankara.

J. Fry, H. Asoh, and T. Matsui. 1998. Natural dialogue with the Jijo-2 office robot. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*.

J. M. Iverson, O. Capirci, E. Longobardi, and M. C. Caselli. 1999. Gesturing in mother-child interactions. *Cognitive Develpment*, 14(1):57–75.

T. Kanda, H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu. 2002. Development and evaluation of an interactive humanoid robot "robovie". In *Proc. Int. Conference on Robotics and Automation*.

W. Levelt. 1989. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

H. J. Levesque, P. R. Cohen, and J. H. T. Nunnes. 1990. On acting together. In *Proc. Nat. Conf. on Artificial Intelligence (AAAI)*.

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*.

D. McNeill. 1992. *Hand and Mind: What Gesture Reveal about Thought*. University of Chicago Press.

M. F. McTear. 2002. Spoken dialogue technology: enabling the conversational interface. *ACM Computing Surveys*, 34(1).

Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. 2003. Towards a model of face-to-face grounding. In *Proc. Annual Meeting of the Association for Computational Linguistics*.

L. J. M. Rothkrantz, P. Wiggers, F. Flippo, D. Woei-A-Jin, and R. J. van Vark. 2004. Multimodal dialog management. In *Proc. Text, Speech, and Dialogues*.

E. A. Schegloff and H. Sacks. 1973. Opening up closings. *Semiotica*, pages 289–327.

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

Do you like this robot? The role of robot behavior, robot personality and user personality

Britta Wrede
Applied Computer Science Group
Bielefeld University
Germany
++49 521 106 2934
bwrede@techfak.uni-bielefeld.de

Stephan Buschkämper
Department of Psychology
Bielefeld University
Germany
++49 521 106 2934
sbuschkaemper@uni-bielefeld.de

Shuyin Li
Applied Computer Science Group
Bielefeld University
Germany
++49 521 106 2952
shuyinli@techfak.uni-bielefeld.de

ABSTRACT
We analyzed if users are able to assign personality traits to a robot and which factors influence liking of the robot. It turned out that users had no difficulties in judging the robot's personality. The analysis of the ratings revealed that the robot's dialog behavior (basic vs. verbose) had an effect on the personality ratings of the robot. Furthermore, the robot's behavior, aspects of the robot's personality, and the user's personality contributed independent proportions of variance in explaining liking of the robot. These results indicate that the personality of users as well as of robots should be taken into account when designing human-robot interfaces.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: Natural language; I.2.9 [**Robotics**]: Operator interfaces

General Terms
User evaluations, Human-Robot dialog, HRI foundations, HRI applications

Keywords
Personality ratings, speech-based human-robot interaction

1. INTRODUCTION
If users are to accept robots in their private lives robots need to blend in the social situation and act according to social rules. One factor that would foster such social blending in (or *social embedding* [1]), is to design robots in such a way that people perceive them as likeable personalities. However, is it possible that a user attributes human personality traits to an artificial system? This seems plausible as not only children but also adults tend to anthropomorphize inanimate objects. Second, it has already been shown that users are able to use descriptions of human personality traits to describe the personality of robots [2]. But do users really feel that such personality descriptions are adequate and does robot personality indeed influence how users feel about the robot? In our study we addressed the following three questions: (1) If asked to describe a robot's personality with traits established in personality psychology, how easy do users find this task and how sure are they about their judgment? (2) Does the robot's behavior influence the perceived personality? (3) Is the robot's personality relevant for whether or not people like the robot and what other factors play a role?

2. DATA COLLECTION
The basis of our data collection was a user study carried out with our mobile robot BIRON (Bielefeld Robot Companion) an interactive robotic system based on an ActiveMedia PeopleBot platform. A basic component of the robot is a person attention system [3] which enables the robot to focus its attention on one person. Based on this attention the robot can physically follow the person of interest and engage in verbal interactions. A multi-modal object attention module allows the system to learn new objects that the user is showing. For the purpose of this study we disabled BIRON's mobility so that it remained fixed on its place.

In total 14 users aged between 25 and 37 interacted with BIRON in two subsequent sessions. The second session was preceded by a technical information explaining the underlying functionality of BIRON in more detail. After the second session the users completed a set of questionnaires regarding their judgment of the interaction as well as their own personality and the perceived personality of the robot. The personality of the robot and the user were each assessed by a time-economic questionnaire, the BFI-10 [4], which measures personality according to the widely accepted and cross culturally applicable Big Five Model of personality [5].

The mean interaction time of each session was 5 minutes. In order to assess the influence of different behavior on the perceived personality of the robot we used two different interaction types distributed randomly over the subjects. The basic interaction type gives only feedback when the robot is addressed by the user. In contrast, the verbose interaction type will actively engage in a conversation by initiating an interaction when the system detects a person and will also give comments relating to the success of the communication at certain points during the interaction (e.g. "It's really fun doing interaction with you" or "I know it's sometimes difficult with me. But please don't feel discourage.").

3. ANALYSIS
After rating the robot's personality users were asked how easy the task of judging BIRON's personality was and how sure they felt about their judgment. 71.4% of all our users felt very or rather sure about their judgment and 57.1% thought the task was

very or rather easy. For both items only 14.3% of our users thought the task was difficult or were not sure about their judgment. Both ratings were substantially correlated (r =.71, p <.01).

One of our main findings was that users interacting with the verbose version of BIRON rated the robot significantly more *extraverted* than users interacting with the less talkative version (*t*-test for independent samples: $p < .05$, see Fig. 1). Interestingly, the standard deviations indicate that the verbose version might also provoke more heterogeneous judgments.

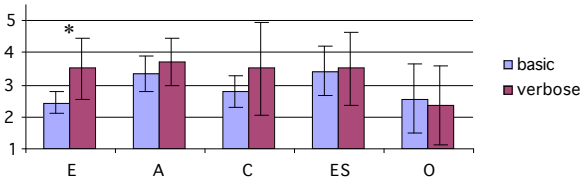

Figure 1: Personality profiles of BIRON with basic vs. verbose dialog behavior (E: Extraversion, A: Agreeableness, C: Conscientiousness, ES: Emotional Stability, O: Openness).

Given that users found it so easy to assign a personality to the robot we wondered if the perceived personality had an influence on whether or not the users liked BIRON or if other factors also played a role. As candidate factors that might have an influence we considered the robot's behavior (verbose vs. basic), the robot personality as rated by the user, and the user's own personality.

By using stepwise multiple regression analysis we identified in a first step the robot's perceived *agreeableness* and the user's *openness to experience* as personality traits that might be useful in explaining variances in users liking BIRON. In a second step these personality traits together with robot behavior (verbose vs. basic) were used as predictors in a multiple regression analysis with simultaneous entry of the predictor variables. The results showed that with these three predictors it was possible to explain 78.1% of the variance of the liking judgments (see Fig. 2). Each of the predictors contributed independently between 11.8% and 22.8% of variance. Note that each of these unique contributions is not explainable by the other two predictors.

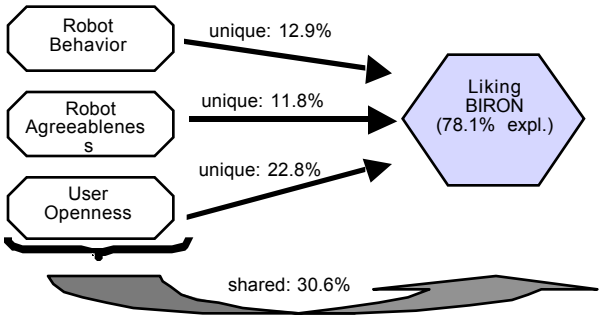

Figure 2: Factors explaining variance in the users' ratings for liking BIRON.

Interestingly, in the course of this analysis it became obvious that the perceived *extraversion* of BIRON is not significantly linked to how much the users like BIRON ($r = -.13$, *n.s.*). Therefore one can conclude that different aspects of BIRON's behavior influence users' *extraversion* rating and the users' liking of BIRON. On the other hand the perceived *agreeableness* of BIRON was not significantly influenced by BIRON's interactive style ($r = .28$, *n.s.*, see also Fig. 1).

4. CONCLUSION

Our data support the hypothesis that attributing human-like personality traits to a socially embedded interacting robot comes naturally to users. Our subjects neither found it difficult to describe the robot's personality nor did they feel uncertain about their judgment.

A second finding was that the robot behavior can influence the users' judgments of the robot's personality traits and possibly also the homogeneity of these judgments. That a significant difference was found only for one out of five personality traits might be due to the nature of the behavior differences, the interaction task, the influence other factors have on the personality rating (e.g. design features or morphology), or any interaction of these aspects. It will be necessary to challenge these results with more data.

Finally, the behavior of the robot, the personality of the robot as judged by the users, and also the user's personality contribute independent proportions of variance to explaining differences in liking the robot. Personality ratings – of the robot as well as of the user - are thus not redundant with behavioral and interaction variables when explaining users' acceptance of a robot.

Personality characteristics of the robot as well as of the users should be taken into account when developing robots for social tasks and for different user groups.

5. REFERENCES

[1] K. Dautenhahn, B. Ogden & T. Quick. From embodied to socially embedded agents – Implications for interaction-aware robots. In: *Cognitive Systems Research, 3(3), 2002*, 397-428.

[2] S. Woods, K. Dautenhahn, C. Kaouri, R. te Boekhorst , K. L. Koay. Is this robot like me? Links between Human and robot personality traits. Accepted for *IEEE Humanoids,* 2005.

[3] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G.A. Fink & G. Sagerer. Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems, 43(2-3), 2003*, 133-147.

[4] B. Rammstedt & O. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Submitted manuscript.

[5] R.R. McCrae & O. John. Introduction to the five-factor model and its applications. *Journal of Personality, 60, 1992*, 175-215.