



FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D1.2006

Multi-modal dialogues

Due date of deliverable: 31/12/2006

Actual submission date: 31/01/2007

Start date of project: January 1st, 2004

Duration: 50 months

Contributing partners:

UniBi, KTH, UH, LAAS

Revision: final

Dissemination Level: PU

Executive Summary

In the third phase of the project the overall objective was to increase the situation awareness of the dialogue.

Based on results from the corpus analyses in WP1.3 as provided by KTH we pursued two strategies: in WP1.1 we concentrated on adapting to different types of users based on evidence from WP1.3 and literature, thus increasing the awareness with respect to the user. In contrast, in WP1.2 we focussed on multi-modal information as derived from different perception components in order to increase the awareness with respect to the interaction situation. This work was carried out by UniBi.

Further steps towards a more intelligent situation awareness have been taken by developing a preliminary conceptual integration of the dialogue with the planning (or supervision) module as provided by RA6. This work has been carried out in cooperation between the partners LAAS and UniBi. Furthermore, we continued our work on spatial representation by integrating the KTH localisation module with the dialogue.

Finally, in WP1.3 we have focussed our interest towards the planning and preparation of evaluation activities for the fourth phase. We have provided input to the evaluation plan by giving a scientific background to the term “user experience” as a basis for the evaluation activities of the key experiments. We have further refined the multi-modal corpus by providing the possibility to utilize context-providing resources (task sequences, positioning, text descriptions etc). This work has been carried out by KTH.

With respect to overall evaluation, video trials of the home-tour scenario in cooperation between the three partners UH, KTH and UniBi have been initiated.

Role of multi-modal dialogues 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

While the dialogue has so far been only integrated in the Home-Tour Scenario (KE1), it is now being integrated in the Curious-Robot Scenario (KE2) in order to allow for an integration of dialogue with task planning.

1	Multi-modal Dialogues	4
1.1	Adaptive multi-modal dialogue	4
1.1.1	Adaptation model	4
1.1.2	Communicative Success	6
1.2	Representation and integration of knowledge for an embodied multi-modal dialogue system	6
1.2.1	Handling of multiple-users situations	6
1.2.2	Integration of planning with dialog modeling	7
1.2.3	Integration of dynamic topic tracking with dialog modeling	8
1.2.4	Integration of localisation (KTH implementation) with dialog modeling	8
1.3	Evaluation methods	8
1.3.1	Using the corpus	8
1.3.2	Work on user types	10
1.3.3	Video Trials	10
2	Future Work	11
3	References	11
3.1	Applicable documents	11
3.2	Reference documents	12
4	Annexes	13
4.1	Summaries of Papers	13
4.2	Papers	18
	Paper [1]	18
	Paper [2]	22
	Paper [3]	30
	Paper [4]	36
	Paper [5]	44
	Paper [6]	50
	Paper [7]	52
	Paper [8]	59
	Paper [9]	67
	Paper [10]	73
	Paper [11]	80
	Paper [12]	84
	Paper [13]	88
	Paper [14]	97
	Paper [15]	100
	Paper [16]	114

1 Multi-modal Dialogues

1.1 Adaptive multi-modal dialogue

1.1.1 Adaptation model

The main focus of WP1.1 in this phase was the design and the implementation of adaptive dialogue strategies based on an appropriate user model. Our approach [15] was motivated by analyses of user types and strategies carried out by KTH in WP1.3 and a large body of literature on adaptive systems in HCI and by our own empirical knowledge in HRI. The goal of this effort is to increase usability and acceptability of the system to different types of users.

Our approach follows the well established general adaptation schema (Structure (a) in Fig. 1) proposed by Jameson [22]. The central construct in this schema is the user model that specifies characteristics, needs, preferences and goals of users that are typical for an application. The content of the user model is often defined in terms variables, the values of which differ between users. During the interaction, the system adapts its behaviors based on the values of the current user. To achieve this goal, the system needs to first collect information about the user based on her input. Then, in the process of user model acquisition, statistical learning algorithms and/or rule-based inference mechanisms are applied in order to associate the information collected from the user with certain variables in the user model. Subsequently, the system applies these relevant values to the on-going interaction and varies its behaviors in certain interaction situations. The challenges in the development of user adaptive systems based on this schema lie in the identification of appropriate user model variables, the association of user input with these variables and the specification of interaction situations, for which the system behaviors should be adaptive.

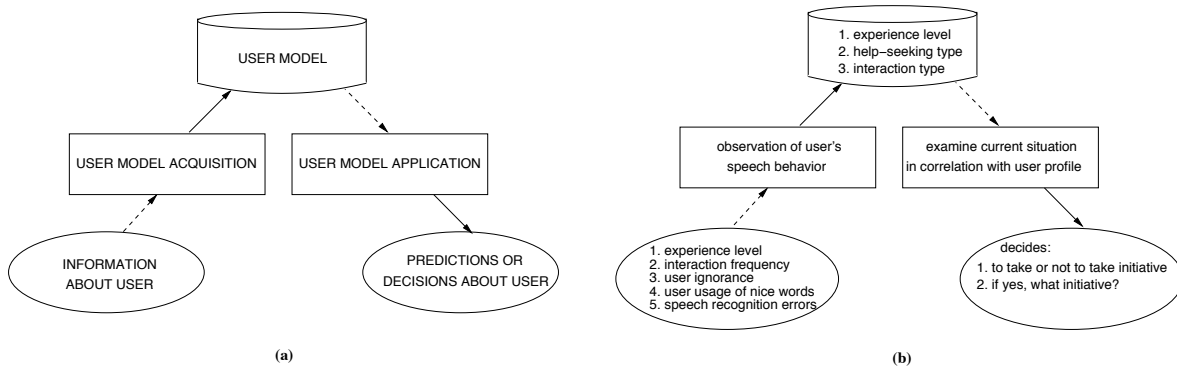


Figure 1: (a) General schema for the processing in a user-adaptive system according to Jameson. (b) The implemented adaptive schema in RA1 (Ovals: input or output; rectangles: processing methods; cylinder: stored information. Dotted arrows: use of information; solid arrows: production of results)

The adaptation mechanism that was implemented within RA1 is based on this general schema (structure (b) in Fig. 1). The major innovation of our approach is the broad coverage of the user model and the system's ability to vary its initiative behaviors.

The main components of the adaptation scheme are the user model, the information collection for the acquisition of the user model and its application in order to produce predictions about the user's behavior as a basis for decisions of the robot's own behavior. These components are summarised in the following.

- **User model:** The user model specifies characteristics of users concerning their interaction expe-

rience and preference. In our current implementation, each user is modeled by three variables: *experience level* (and interaction frequency), *help-seeking type* and *interaction type*, which address the issue of task success, interaction control and interaction pleasure, respectively.

- The *User experience level* is a heuristic measure for the user's initial need for task-related help from the system. The values can either be *novice* or *experienced*. *Interaction frequency* is a more fine-grained variable and enables even finer variation of system behaviors.
 - The *help-seeking type* of a user indicates the attitude of the user with regard to her preference for the system's task-related initiative. In this respect, a user can be *help-appreciator* who is grateful for hints about how to interact with the system, or she can be *help-rejector* who prefers to explore the system by herself and views system's suggestions as an attempt to patronize her. For example, a novice user can be expected to need more task-related help from the system. However, if she is a help-rejector the system should phrase its help suggestions more carefully and less frequently.
 - The *interaction type* can be interpreted as an attempt to estimate the user's tendency to anthropomorphize the robot. According to Fischer [19], a user can be a *player*, who treats a robot as a human and enjoys socially-aware behavior of the robot, or a *non-player*, for whom a robot is a machine whose performance is measured only according to its successful task execution rate and who should not make socially-related comments.
- **Information collection:** The above variables are instantiated based on the user's interaction behaviors. More specifically, the following information from the user is collected: (1) the interaction frequency of a given user (for user *experience level* and user *interaction frequency*); (2) whether the user follows system's task-related initiatives, that is, whether she reacts positively to help given by the system (for *help-seeking type*); (3) whether the user uses phrases that are typical for human-human interactions, e.g., "Thank you", "Sorry", "How are you?" (for *interaction type*). Additionally, the amount of speech recognition errors that occur after utterances that are initiated by the system is registered. This is a useful measure for estimating the user's tendency of meta-commenting the system's behavior so that steps can be taken to reduce the damage of some users' meta-comments relating to the system's behavior.
 - **User model acquisition:** Once a user is identified at the beginning of the interaction, her values for experience level and interaction frequency are increased. Once a user demonstrates the above behaviors (2) and (3), her values concerning help-seeking type and interaction-type are updated. Note, there is no one-to-one causal relationship between behaviors and values for the corresponding variables, instead, the behaviors (2) and (3) have to occur more frequently than a threshold until the user model values are actually changed. This policy softens the behavior variations of the system.
 - **User model Application:** During an interaction, the dialog system interpretes interaction situations (s. below) depending on the values of the user model. Some of the interaction situations require an online evaluation of the user's and the system's performance. An indicator for performance problems of the user are the number of illegally proposed tasks. Indicators for robot performance problems are perception problems such as speech and gesture recognition errors.
 - **Prediction or decision about the user:** The behavior variations of the system concern *the content* and *the frequency of the system's initiatives*. Four attributes are specified for each initiative:

Type (task-related or social initiatives), *Group* (min, max, or med, indicating the frequency of initiative that should be used for the current user), *Trigger* (the interaction situation, in which the system has to make a decision about taking the initiative) and *Priority* (priority of this initiative, relevant in case of multiple possible initiatives). Given a trigger and the current user profile (based on the values of the user model), the system makes a decision on whether to take an initiative of either type and/or what initiative to take. These initiatives are defined in an external configuration file and include initiatives such as proposing self-explanation at the beginning of the interaction (task-related initiative), proposing new legal task (task-related initiative), showing happiness when starting interaction with a known user (social initiative), summarizing the current session at the end of the interaction (social initiative), and so on. For a complete list of these initiatives and how they are handled given different user profiles, see [15].

Our adaptation approach is a powerful mechanism and its benefits include the following three points. Firstly, the three variables of the user model address both aspects of *usefulness* and *social ability* of a robot companion (as identified as crucial by Dautenhahn [18],) and enable them to be operationalized with clearly defined system behaviors. Secondly, the parameters of initiative frequency softens the adaptation behavior of the system and reduces user confusion caused by variable system behavior (which is a common problem for adaptive systems). Thirdly, the system can be implemented in a way that the major behavior control is sourced out to configuration files so that modifications can be easily carried out. For the details of the implementation, including the currently implemented initiatives, triggers, update mechanism and the major configuration files, see [15].

1.1.2 Communicative Success

Our work on communicative success has been integrated implicitly in the adaptation model. In this model we operationalize communicative success by observing the number of illegal commands given by the user and by the (perceived) recognition errors of the system. From this perspective, our goal of the adaptation scheme is to enhance the communicative success by minimizing the ‘errors’ (i.e. unexpected behaviors) made by the user. In parallel to this adaptation scheme we still maintain the concept of support exchanges (= clarification subdialogues). The frequency of support exchanges is used as an online measurement of the performance of the dialogue system.

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

The focus of this workpackage was to increase the situation awareness of the dialog system by taking into account multi-modal information about the interaction situation.

1.2.1 Handling of multiple-users situations

We implemented a mechanism for the system to be able to correctly make the decision whether to process speech input forwarded by the speech recognizer, which is important for the robot to achieve situation awareness. The necessity for this mechanism is demonstrated in the following two situations. In situation one, the user is talking to the robot while another person approaches the user to talk with her. The major challenge of this situation is that it is not clear whether the dialog system should process the speech input forwarded by the speech recognizer because it can not be determined for sure, by the perception modules alone, whether the user speech is now directed to the robot or to the third person. In situation two, the robot is driving into the kitchen autonomously and the user suddenly

shouts “stop”. Usually, this speech input will not be processed because the user is behind the robot and no active interaction partner can be perceived. This “stop” can be, however, crucial when the robot is on the wrong track and in danger of colliding with an object.

To cope with such situations, the dialog system draws on three different parameters to make this decision: the content of the speech input, the input from the person detection module and the dialog expectation. As can be seen in the pseudo-code in Fig. 2, the more ambiguous the situation is, the more restrictions are put on the decision making process. As an emergency command “stop” is allowed in any situation. Even if the “stop” is not directed towards the robot, the cost of this wrong decision is probably lower than vice versa. If a human can be detected, then the decision depends on whether she looks at the robot as she produces speech. If more than one person are detected, the decision mainly relies on the dialog system’s expectation: if a user reply is expected, then the speech is probably directed to the robot, otherwise, it again depends on whether the user is looking at the robot. In case the situation is classified as ‘ambiguous’ error handling strategies (*handle-suboptimal-case*) are applied depending on the user adaptation process (i.e. by explicitly asking for clarification or by ignoring ambiguous command). The graded decision making process of this approach has the advantage that different levels of restrictions can be adopted in different situations.

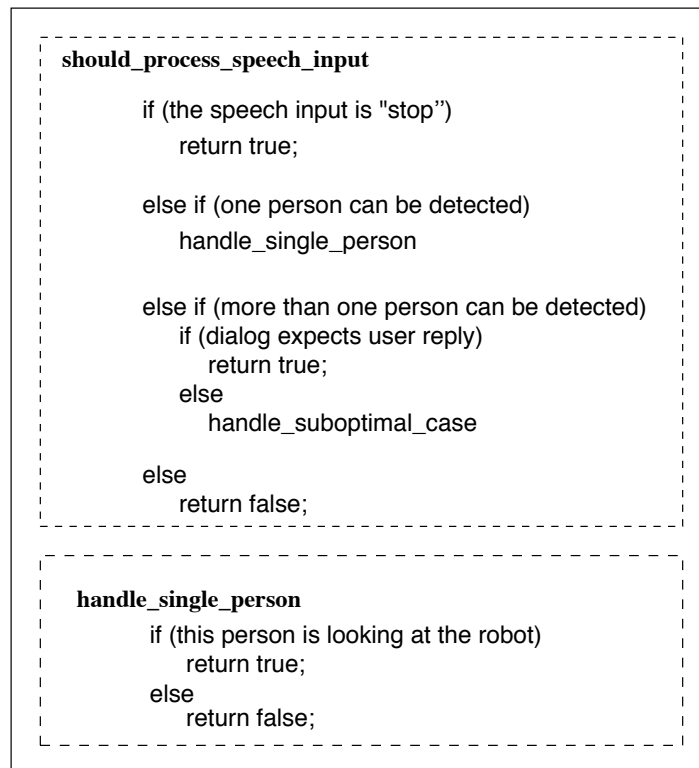


Figure 2: How to decide whether to process speech input

1.2.2 Integration of planning with dialog modeling

Apart from sensoric information that gives information about the interaction situation, situation awareness can also be improved by incorporating reasoning mechanisms that are provided by a supervision or planning component. Therefore, a conceptual integration of dialog and task planning has been

prepared in close cooperation with RA6. The scenario, which will be integrated in KE2, envisions currently the task of asking the robot to fetch a certain object from a certain place. While this is an apparently trivial task with respect to dialog it has turned out to consist of numerous interesting decision points where either the user or the robot can take initiative in order to negotiate about the current plan (for an internal working paper on this work cf. [13]).

1.2.3 Integration of dynamic topic tracking with dialog modeling

In order to demonstrate situation awareness with respect to the semantics of the user's utterances we integrated an approach to dynamically track topics in situated human-robot interactions. By dynamically we mean that neither the topics nor the amount of topics are predefined but adjusted dynamically during the interaction. This enables the system to draw simple inferences without explicit reasoning and without a pre-defined knowledge-base by associating words and objects that 'belong', that is co-occur together.

The topic tracker has been integrated with the dialogue module and has been evaluated offline (precision and recall of topic classification) and online. It was shown that subjects were sensitive to the robot's ability to detect connections in the discourse ([10], [11] [12]).

1.2.4 Integration of localisation (KTH implementation) with dialog modeling

Furthermore, we continued our work on spatial representation by integrating the localisation approach from KTH with the dialogue. Based on experiences from user studies in RA1 and RA3 a mapping hierarchy was defined where spatial points are represented as either *rooms*, *areas* or *objects*, depending on how users refer to them and what role they play in further interactions. Our hypothesis is that it will be easier for users to interact with a robot when they have similar representations of the spatial environment. See [16] for an unpublished detailed summary of this work.

Further publications within this WP in the third phase concern results from work accomplished mainly in the second phase of the project [9], [6], [8], [7], [4] and [5].

1.3 Evaluation methods

In the third phase we have focussed our interest towards the planning and preparation of evaluation activities for the fourth phase. We have provided input to the evaluation plan by giving a scientific background to the term "user experience" as a basis for the evaluation activities of the key experiments.

We have further refined the multimodal corpus by providing the possibility to utilize context-providing resources (task sequences, positioning, text descriptions etc). In coordination with work in WP3.4 we have refined the annotation schema based on the Cogniron taxonomy. This has been reported in [3] and in the deliverable for RA3 (WP3.4).

1.3.1 Using the corpus

The corpus has been utilized to propose and empirically motivate the novel concept of *spatial prompting*¹ that can be used to actively influence users' way of positioning themselves. This is further described in [1].

¹A *spatial prompt* can be defined as *the act of inciting someone to perform spatial action*.

By viewing the video corpus we identified and analyzed instances where the robot movements or verbal actions appear to influence the actions of the user. The robot seems to incite the user to act in different ways:

- Primary verbal: by using a spoken command
- Primary non-verbal: by movements
- Multi-modal: using movement as trigger for a verbally specified (or grounded action)

An example of a spatial prompt for a cognitive robot would be to direct the camera gaze towards the direction the robot is going in order to implicitly “give way” to the user.

We have also used the corpus data and the available literature to analyse generic user behavior in a human-robot interaction context. We have approached this from two perspectives. First of all we have attempted to find support in our corpus data for previously reported user behavior or user classification (e.g., *player* vs. *non-player* users [19], *degree of engagement* [24]).

The second perspective is more explorative, but from a project perspective more interesting. To our knowledge, very few attempts to model a user scenario as complex as the Home tour exist. We have therefore also categorised users, trying to group them into broad classes based on behaviors observed in the corpus.

This work is summarised in an internal working paper [14] and is also heavily related to the work on development of a joint taxonomy of gesture for human-robot interaction done in RA3 (WP3.4) as the annotation categories are used to derive, explain and motivate the user classification. If we are to use behavior of users as cues for classification we need to establish to what extent this is done using

- Human observation, i.e., through qualitative analysis of usage situations (e.g. corpus analysis)
- Algorithmic methods, i.e, by employing robot sensors and software to pick up and classify user behavior.

Since there is a link (through the use of the taxonomy) to the user classifications in the corpus, we are also taking into consideration to what degree algorithmic methods can be employed to perform classification of users in the real system. Currently in WP1.3 we are analysing the corpus (i.e. using human observation) to come up with a set of user behaviours that can be used to define a broad user model. We are thus analysing user behavior along three dimensions:

- Global organisation: interaction and task strategies. In RA3 we have analyzed how task transitions affect the way spatial positioning is performed [21]. The data from this study provides a global picture of the distribution and sequencing of tasks for the Home Tour scenario. Furthermore we have been looking at the following episodes in the corpus data:
 - First encounter: how is interaction initiated?
 - First failure: when is the first problem encountered?
 - Organisation of task: what are the interaction characteristics with respect to solving the task?
 - Closing: how is the interaction ended?
- Local organisation: we are analysing sequences of user behavior in order to establish single or complex behaviors that can be observed by humans and possibly used as cues for algorithmic methods. One example is the way users sometimes change their posture (e.g. folding arms)

after issuing a “find” command. Another example is how objects are manipulated in order to provide for a show interaction episode.

- Natural language adaptations: we are analysing the corpus to establish if there are linguistic and gestural patterns that can provide cues for modelling users. For instance, the way users align to the robot’s vocabulary; how pointing gestures are performed; or the way some users attempt to establish eye-contact with the robot’s onboard camera.

Work in this workpackage has been published in [1], [2], [3].

1.3.2 Work on user types

Based on analyses of the corpus we have come up with the proposal of two different types of information that could be used to group users based on the very first robot encounter:

- Greeting, first actions and first problem encountered,
- Distribution of deictic and ergotic (manipulative) gesture.

The first type of information can be accounted for by using, e.g., an information-state approach used in dialogue systems. The second type, provides a challenge for a sensor based system. Thus in order to encounter for this type of activities we need to detect both pointing gestures and ergotic gestures. It is therefore necessary to allow for any component that maintains a user model to be able to receive or get information from other system components and should be considered in system design. Since this work is explorative we will turn our future interest towards behavioural features of interaction that could be used to tie into other frameworks, for instance, the notion of engagement [17, 24]. And also, with respect to Fischer’s notion about the *player/non-player* [20], the notion of *presence as immersion* [23] and *willingness to suspend disbelief* [23]. For more details of this analysis cf. to the unpublished working paper [14].

1.3.3 Video Trials

As a pre-study for the evaluation of the dialogue module in real life interactions we are currently carrying out a video study as a joint effort of UH, KTH and UniBi. For this study we have video-taped an interaction of an actress with BIRON in the apartment that has been rented by UniBi. Alternative behaviours of BIRON with respect to the dialogue behaviour and the spatial movements² have been recorded in order to compare them by assessing the ratings of subjects watching the video-taped human-robot interaction.

The video and questionnaires are currently being presented in a pre-trial to students at UniBi. For the final study the video and questionnaire will be translated to English and Swedish in order to draw on a larger population and to allow for inter-cultural comparisons.

²Note that for the ease of recording we have completely teleoperated BIRON for this study. For some behaviours this was even necessary as not all functionalities were completely and robustly integrated at the time of recording (November 2006).

2 Future Work

In the final phase we will focus our efforts on the integration of dialogue and task planning by fully integrating the dialogue system in the KE2 platform. The second focus in RA1 will be on the evaluation of the dialogue as within the overall system by real interaction studies. An extension of our work will be user studies within the home tour scenario (KE1) with respect to user experience of the overall system when interacting with the final KE1 platform.

3 References

3.1 Applicable documents

Published

- [1] Anders Green and Helge Hüttenrauch. Making a case for spatial prompting in human-robot communication , multimodal corpora: From multimodal behaviour theories to usable models. In *Workshop at the Fifth international conference on Language Resources and Evaluation (LREC)*, Genova, 2006.
- [2] Anders Green, Helge Hüttenrauch, and Elin Anna Topp. Measuring up as an intelligent robot - on the use of high-fidelity simulations for human-robot interaction research. In *Performance Metrics for Intelligent Systems, The 2006 Performance Metrics for Intelligent Systems Workshop (PerMIS)*, Gaithersburg, MD, 2006.
- [3] Anders Green, Helge Hüttenrauch, Elin Anna Topp, and Kerstin Severinson Eklundh. Developing a contextualized multimodal corpus for human-robot interaction. In *Fifth international conference on Language Resources and Evaluation (LREC)*, Genova, 2006.
- [4] Sonja Hüwel and Britta Wrede. Situated speech understanding for robust multi-modal human-robot communication. In *Proceedings of the International Conference on Computational Linguistics (COLING/ACL)*. ACL Press, 2006.
- [5] Sonja Hüwel, Britta Wrede, and Gerhard Sagerer. Robust speech understanding for multi-modal human-robot communication. In *Proc. 15th Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE Press, 2006.
- [6] Shuyin Li. Position paper. In *Proc. 2nd Young Researchers' Roundtable on spoken dialog systems, In conjunction with Interspeech 2006*, pages 43–44, 2006.
- [7] Shuyin Li and Britta Wrede. Why and how to model multi-modal interaction for a mobile robot companion. In *Proc. AAI Spring Symposium on Interaction Challenges for Intelligent Assistants*, Stanford, 2007. to appear.
- [8] Shuyin Li, Britta Wrede, and Gerhard Sagerer. A computational model of multi-modal grounding. In *Proc. ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006*, pages 153–160. ACL Press, 2006.
- [9] Shuyin Li, Britta Wrede, and Gerhard Sagerer. A dialog system for comparative user studies on robot verbal behavior. In *Proc. 15th Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 129–134. IEEE Press, 2006.

- [10] Jan F. Maas, Thorsten Spexard, Jannik Fritsch, Britta Wrede, and Gerhard Sagerer. BIRON, what's the topic? - A Multi-Modal Topic Tracker for improved Human-Robot Interaction. In *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, pages 26–32. IEEE, September 2006.
- [11] Jan F. Maas and Britta Wrede. BITT: A corpus for topic tracking evaluation on multimodal human-robot-interaction. In *Proceedings of the international conference on Language and Evaluation (LREC)*, Genoa, Italy, 2006.
- [12] Jan Frederik Maas, Britta Wrede, and Gerhard Sagerer. Towards a multimodal topic tracking system for a mobile robot. In *Proc. International Conference on Spoken Language Processing (Interspeech/ICSLP)*, pages 1189–1192, September 2006.

Internal Papers and Technical Reports

- [13] Aurélie Clodic, Shuyin Li, Britta Wrede, and Rachid Alami. Interaction between dialog, supervision and planning, Joint RA1-RA6 working paper, UniBi-LAAS. Technical report.
- [14] Anders Green. User Models Based on Initial Interaction Patterns and Deictic / Ergotic Gestures - Working paper on user models for adaptation, KTH. Technical report.
- [15] Shuyin Li and Britta Wrede. Adaptive dialogue for human-robot interaction. Technical report, Applied Computer Science, Bielefeld University, 2006.
- [16] Thorsten Spexard, Elin A. Topp, Helge Hüttenrauch, Britta Wrede, Shuyin Li, and Gerhard Sagerer. Working paper on the integration of a Human Oriented Map Representation with a Grounding Based Dialogue Module - Summary of ongoing work on localisation and dialogue integration, 2006. Technical report.

3.2 Reference documents

Other References

- [17] Cory D. Kidd Neal Lesh Candace L. Sidner, Christopher Lee and Charles Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166:140–164, 2005.
- [18] K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K.L. Koay, and I. Werry. What is a robot companion - friend, assistant or butler? In *Proc. IEEE IRS/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1488–1493, Edmonton, Alberta Canada, 2005.
- [19] K. Fischer. *What Computer Talk Is and Is not: Human-Computer Conversation as Intercultural Communication*, volume 17 of *Linguistics - Computational Linguistics*. AQ, Saarbrücken, 2006.
- [20] Kerstin Fischer. The role of users' preconceptions in talking to computers and robots. In *Proc. of the Workshop on 'How People Talk to Computers, Robots, and Other Artificial Communication Partners'*, pages 112–130, Delmenhorst, Germany, 2006.
- [21] Helge Hüttenrauch, Kerstin Severinson Eklundh, Anders Green, Elin Anna Topp, and Henrik I. Christensen. What's in the Gap? Interaction Transitions that make HRI work. In *Proceedings*

of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006), Hatfield, UK, Sep 6–8 2006.

- [22] Anthony Jameson. *Human-Computer Interaction Handbook*, chapter Adaptive Interfaces and Agents. Erlbaum, Mahwah, NJ, 2003.
- [23] Matthew Lombard and Theresa Ditton. At the heart of it all: The concept of presence. 1997.
- [24] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. Where to look: a study of human-robot engagement. In *IUT'04: Proceedings of the 9th international conference on Intelligent User Interfaces*, pages 78–84, New York, NY, USA, 2004. ACM Press.

Annexes

The annex contains a set of published and internal papers as listed below in alphabetical order.

Published Papers

[1] Making a Case for Spatial Prompting in Human-Robot Communication , Multimodal Corpora: From Multimodal Behaviour theories to usable models

Anders Green, Helge Hüttenrauch

Workshop at the Fifth international conference on Language Resources and Evaluation, LREC2006, Genova, May 22-27.

This paper motivates a design case for a multimodal interface where the robot utilizes a (naïve) spatial prompting strategy to influence the spatial positioning and communicative behavior of the user. The design can be seen as an example of *spatial prompting*, a notion which is introduced in this paper. To provide an empiric background for this design we have analysed a set of examples of how verbal and non-verbal behavior of a service robot influence users' way of positioning themselves during interaction, using concepts from theories of non-verbal behavior.

[2] Measuring Up as an Intelligent Robot - On the Use of High-Fidelity Simulations for Human-Robot Interaction Research

Anders Green, Helge Hüttenrauch, Elin Anna Topp (2006)

PerMIS '06, Performance Metrics for Intelligent Systems, The 2006 Performance Metrics for Intelligent Systems Workshop, Gaithersburg, MD, August 21-23, 2006

This paper describes and discuss user-centered Hi-Fi simulation methods. By actively shaping the system both in terms of technology and in interface design the interest of both users and designers are involved in the process as stake-holders. We have analyzed two different simulation studies performed sequentially with different foci, but within the same overall scenario. The data obtained from Hi-Fi simulation studies is to a large extent qualitative, but data can also be used as a resource to be used for component development, e.g. as training data for speech recognizers. The collected data may also be used for quantitative evaluation depending on the experiment design.

[3] Developing a Contextualized Multimodal Corpus for Human-Robot Interaction

Anders Green, Helge Hüttenrauch, Elin Anna Topp and Kerstin Severinson Eklundh

Fifth international conference on Language Resources and Evaluation, LREC2006, Genova, May 22-27.

This paper describes the development process for the contextualized corpus that has been developed. The data have been collected in two Wizard-of-Oz user studies performed with 22 and 5 users in the the Home Tour scenario. The corpus has been transcribed and annotated with respect to gestures and conversational acts, thus forming a core annotation. We have also annotated or linked other types of data, e.g., laser range finder readings, positioning analysis, questionnaire data and task descriptions, that form the annotated context of the scenario.

[4] Situated speech understanding for robust multi-modal human-robot communication

Sonja Hüwel and Britta Wrede

Int. Conf. Computational Linguistics (CoLing/ACL) 2006.

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

[5] Robust speech understanding for multi-modal human-robot communication

Sonja Hüwel, Britta Wrede and Gerhard Sagerer

15th Int. Symposium on Robot and Human Interactive Communication (RO-MAN) 2006.

In order to model complex human robot interaction researchers not only have to consider different tasks but also to handle the complex interplay of different modules of one single robot system. In our context we constructed a robot assistant integrated in a home or office environment. We allow for a fairly natural communication style, which means that the users communicate using speech but are also allowed to use gestures and moreover to use contextual scene knowledge. Against this background, this paper presents a robust speech understanding component for situated human-robot communication. It serves as interface between speech recognition and dialog management. To increase robustness of speech processing it rates the speech recognition output by means of semantic coherence. Even if the recognized word-stream is not grammatically correct the speech understanding component provides semantic interpretations in context of multi-modal input for dialog management. For the understanding process, we designed special semantic concepts grounded to the domain of situated communication. They also provide additional information about the dialog act. A processing mechanism uses these concept units to generate the most likely semantic interpretation of the utterances.

[6] Position Paper

Shuyin Li

2nd Young Researchers' Roundtable on spoken dialog systems In conjunction with Interspeech, 2006

Short position paper presented at the Young Researchers' Roundtable as a basis for discussion on spoken dialog systems.

[7] Why and how to model multi-modal interaction Shuyin Li and Britta Wrede

AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants, 2007. To appear

Verbal and non-verbal interaction capabilities for robots are often studied isolated from each other in current research trend because they largely contribute to different aspects of interaction. For a robot companion that needs to be both useful and social, however, these capabilities have to be considered in a unified, complex interaction context. In this paper we present two case studies in such a context that clearly reveal the strengths and limitations of these modalities and advocate their complementary benefits for human-robot interaction. Motivated by this evidence we propose a powerful interaction framework which accounts for flexible multi-modality management in complex interactions.

[8] A computational model of multi-modal grounding

Shuyin Li, Britta Wrede and Gerhard Sagerer

ACL SIGdial Workshop on discourse and dialog In conjunction with CoLing/ACL, 2006

Dialog systems for mobile robots operating in the real world should enable mixed-initiative dialog style, handle multi-modal information involved in the communication and be relatively independent of the domain knowledge. Most dialog systems developed for mobile robots today, however, are often system-oriented and have limited capabilities. We present an agent-based dialog model that are specially designed for human-robot interaction and provide evidence for its efficiency with our implemented system.

[6] A dialog system for comparative studies

Shuyin Li, Britta Wrede and Gerhard Sagerer

15th Int. Symposium on Robot and Human Interactive Communication (RO-MAN), 2006

In domestic social robot systems the dialog system is often the main user interface. The verbal behavior of such a robot, therefore, plays crucial role in human-robot interaction. Comparative user studies on various verbal behaviors of a robot can effectively contribute to human-robot interaction research. In this paper we present a dialog system that can be easily configured to demonstrate different verbal, initiative-taking behaviors for a robot and, thus, can be used as a platform for such comparative user studies. The pilot study we conducted does not only provide strong evidence for this suitability, but also reveals benefits of comparative studies on a real robot in general.

[10] BIRON, whats the topic? - A Multi-Modal Topic Tracker for improved Human-Robot Interaction.

Jan F. Maas, Thorsten Spexard, Jannik Fritsch, Britta Wrede, and Gerhard Sagerer

Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN), 2006.

Creating robots with extendable social skills and interaction capabilities that suffice their operation in the real world with naive users is a very challenging task. In this paper we present a new approach using topic tracking on multi-modal dialogue to provide a mobile robot with a higher level situation awareness in human-robot interaction. The robot is no longer operating in laboratory surroundings, but in its own real world flat. We describe how our topic tracking approach is implemented in this integrated system, operating on verbal speech input. Different modalities like data from video cameras and laser scans are used as additional cues to a semantic understanding and grouping of user utterances into different topics. Both the amount of topics and the according topic names are created dynamically. Evaluating an offline speech corpus demonstrates the suitability of our approach. It is now possible to ask 'BIRON, whats the topic?', making the interaction more social.

[11] BITT: A corpus for topic tracking evaluation on multimodal human-robot-interaction.

Jan F. Maas and Britta Wrede

Proc. Int. Conf. on Language and Evaluation (LREC), 2006.

Our research is concerned with the development of robotic systems which can support people in household environments, such as taking care of elderly people. A central goal of our research consists in creating robot systems which are able to learn and communicate about a given environment without the need of a specially trained user. For the communication with such users it is necessary that the robot is able to communicate multimodally, which especially includes the ability to communicate in natural language. We believe that the ability to communicate naturally in multimodal communication must be supported by the ability to access contextual information, with topical knowledge being an important aspect of this knowledge. Therefore, we currently develop a topic tracking system for situated human-robot communication on our robot systems. This paper describes the BITT (Bielefeld Topic Tracking) corpus which we built in order to develop and evaluate our system. The corpus consists of human-robot communication sequences about a home-like environment, delivering access to the information sources a multimodal topic tracking system requires.

[12] Towards a multimodal topic tracking system for a mobile robot

Jan F. Maas, Britta Wrede, and Gerhard Sagerer

Proc. Int. Conf. on Spoken Language Processing (Interspeech/ICSLP) 2006.

Topics in situated and task oriented communication depend heavily on the given, often changing environment, making the detection of predetermined topics in many cases useless. Detection of non-predefined topics can enhance Human-Robot-Interaction (HRI) in a variety of ways, though. In this paper we propose a way to dynamically determine topics during Human-Robot-Communication using well established techniques such as Latent Semantic Analysis (LSA). The procedure is based on multimodal cues, supporting the view that topics are not simply a property of spoken or written language, but of multimodal situated communication. An online version of the topic detection system has been developed and is currently being tested on our mobile robot BIRON. To demonstrate the feasibility of our approach, we present the results of an evaluation of our system on the BITT corpus.

Unpublished Working Papers and Technical Reports

[13] Interaction between dialog, supervision and planning

Aur lie Clodic, Shuyin Li, Britta Wrede, Rachid Alami

Joint RA1-RA6 working paper, UniBi-LAAS

[14] User Models Based on Initial Interaction Patterns and Deictic / Ergotic Gestures

Anders Green

Working paper on user models for adaptation, KTH

[15] Adaptive dialogue for human-robot interaction

Shuyin Li, Britta Wrede

Technical Report, Applied Computer Science Group, Bielefeld University, 2006

[16] Working paper on the integration of a Human Oriented Map Representation with a Grounding Based Dialogue Module

Thorsten Spexard, Elin A. Topp, Helge Hüttenrauch, Britta Wrede, Shuyin Li, Gerhard Sagerer
Unpublished summary of ongoing work on localisation and dialogue integration

Making a Case for Spatial Prompting in Human-Robot Communication

Anders Green
KTH School of Computer Science
100 44 STOCKHOLM
green@csc.kth.se

Helge Hüttenrauch
KTH School of Computer Science
100 44 STOCKHOLM
hehu@csc.kth.se

ABSTRACT

In this paper we present an analysis of a set of examples of how verbal and non-verbal behavior of a service robot influence users' way of positioning themselves during interaction, using concepts from theories of non-verbal behavior. Based on the analysis we propose a design case where a robot utilizes a (naïve) *spatial prompting* strategy to influence the spatial positioning and communicative behavior of the user.

INTRODUCTION

A design requirement of a personal service robot is that it should be configured and provided with work tasks by the user in an interactive and intuitive way. These robots are intended to provide service tasks in the home, possibly offering wide range of services. Typically they are envisioned to be equipped with multimodal spoken dialogue systems, to reduce the complexity in the user interface.

In this paper we argue that theories of spatial positioning need to be considered when developing the communicative system of the robot. Furthermore we present an empirical account of the way spatial behavior of robots influence human users. We also propose the term *spatial prompting*, which refers to active strategies of the robot that are intended to influence users to position themselves in a way that is advantageous for further communicative actions.

Positioning, as it has been approached as a research challenge for human-robot interaction, is considered as providing adaptive *physical* movements of the robot. A result of this is that the *communicative dimension* of positioning typically has been ignored in systems that interactively position themselves in relation to their users. One requirement that is typically put forward is that the robot should position itself in a *socially appropriate manner* [1, 4]. The parameters that concern these approaches are typically derived from research on non-verbal behavior.

In robotics the problem of maintaining the robot localized and situated within a geometric representation of the world has been framed as the Simultaneous Localization and Mapping (SLAM) problem [13]. Recent advances in

Human-Robot Interaction (HRI) have raised the interest in detecting and tracking the position of users during interaction. When the position of the user is known, the robot can plan how to position itself [1].

The research on spatial reasoning applied to robotics is well advanced but primarily focused on natural language understanding of spatial relations, providing for exchanges concerning locations of objects in the environment [3].

RESEARCH ON SPATIALITY IN COMMUNICATION

There are several research approaches for human-human that are relevant for spatial management between humans and robots. Hall [14] studied interpersonal distances and distinguished four different distances: *intimate* (0-1.5 ft), *personal* (1.5-4 ft), *social* 4-12 ft, and *public* (> 12 ft). These distances vary both with respect to the current activity and cultural factors. Another dimension that is relevant to spatiality is the concept of *territoriality*, according to Sack, i.e., “the attempt by an individual or group to affect, influence, or control people, phenomena, and relationships, by delimiting and asserting control over a geographic area” [15].

Kendon [12] also studied the spatial configuration of the participants, using the term F-formations, for instance the *L-shape* which describes the relation when two participants have a common visual focus. The shared space, the so called *o-space*, or the *transactional space* is then located in front of the participants, and it is within this area that the interaction is conducted. Clark [5] refers to this space as the *workspace*, where perceptual co-presence is established between speakers [5, 10]. In this context, research on perception and especially visual perception plays an important part for maintaining common ground between participants [10, 8]. Gill [9] has investigated the communicative effects that participants achieve by using nonverbal behavior, focusing on the functional rather than the morphological perspective of nonverbal behavior. One such function is the category *focus* which is a meta-discursive function that signals a shift in the center of attention in the discussion, e.g., a shift in body posture with the same meaning as the utterance “I am going to focus on this spot”.

Another, less obvious, but nevertheless important concept is Schegloff's notion of *body torque* [15], a state of the bodily configuration when two different body segments are oriented in different directions. According to Schegloff [15] Body torque "project change", i.e., when some part of the body is organized in an unstable way, the participants may predict that a change in posture is pending. For instance, when turning the head, this might predict a change of the general body orientation. During interaction, speakers monitor the action of others, interpreting purposeful actions that lead towards a common joint goal as compliance [10]. Human-robot interaction is situated in a physical context, where understanding and reference to actions of the human partner during interaction explicitly needs to be taken into account. This makes research on *virtual collaborative* [6] environments interesting also in this context, since it is concerned with models that explicitly represent spatiality and reference.

CORPUS ANALYSIS OF SPATIAL MANAGEMENT

We have analyzed a video corpus, collected in a European project [7], containing transcribed data of about 20 user sessions, (approximately 20 minutes each) where a user talks to a robot and teaches it the names and locations of objects using a combination of gestures and speech.

By viewing the video corpus we identified and analyzed instances where the robot movements or verbal actions appear to influence the actions of the user. The examples reflect three different ways in which the robot actively influences the user to act:

- Primary verbal: by using a spoken command
- Primary non-verbal: by movements
- Multi-modal: using movement as trigger for a verbally specified (or grounded action)

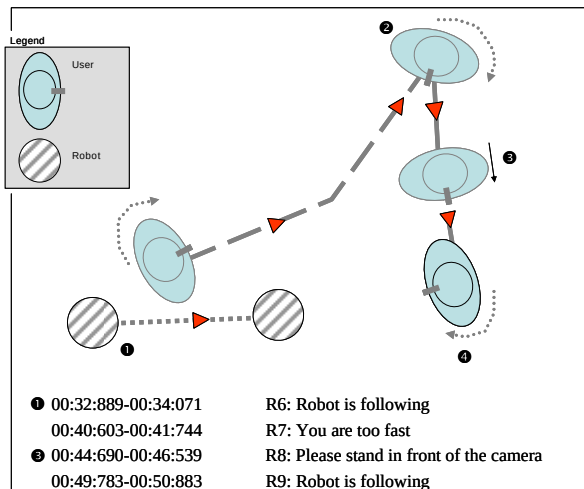


Figure 1: The user and the robot in a follow sequence. The dotted lines show the trajectory of the robot and user.

Primary verbal influence (Example 1)

The example in Figure 1 describes how the events unfold as the user has commanded the robot to follow, after

acknowledging this in (①) the robot starts moving. The robot follows and the user follows the paths depicted in Figure 1. In the second phase of the follow sequence the user has moved to a position that the robot considers too far away. Then the robot says "You are too fast", which triggers the behavior in (②): the user turns towards what can be characterized as the robot's transactional space (according to Kendon [12]) or workspace (according to Clark [5]). Then the user starts moving slightly towards the robot workspace. When the robot speaks "Please stand in front of the camera", the user quickly moves in front of the robot (③-④), something which may be seen as a Give-turn Body Move (according to Gill [9]) that may be seen as a display of the user's willingness to interact [11]. In other words, the user displays her attention towards the robot. The visual attention is aimed at the robot once the user has turned around (throughout phases ②-④).

The first example (Figure 1) illustrates how verbal action directives influence the physical actions of the user. There are instances of this type of example in each of the 20 sessions that the corpus covers.

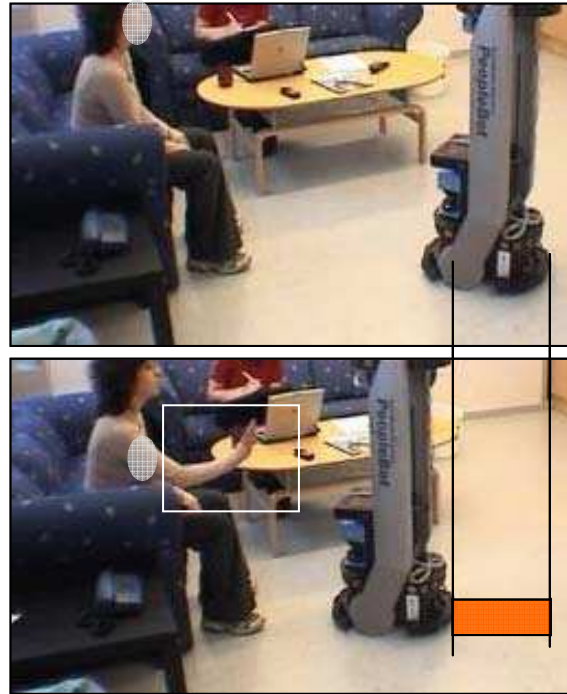


Figure 2: An example of a non-verbal action triggering user action, i.e., a stop gesture (the square in the lower image). The field in the second image, with lines pointing to the first image, illustrates the movement of the robot platform before the gesture is displayed.

Primary non-verbal influence (Example 2)

An example of how the (non-verbal) movements of the robot platform can trigger actions of the user is depicted in Figure 2. In this case the action triggered is a gesture, but it could be another action. In the preceding sequence (not shown) the user has commanded the robot to follow (by

saying “Follow me”). Then the user sits down and waits as the robot is approaching as seen in Figure 2 (upper image). During the approach the user raises the arm and displays a “Stop” gesture. It appears as if the robot comes too close; perhaps crosses the border between a *social* to an *intimate* distance, in terms of Hall [14] or triggers a behavioral reaction as the robot breaches a territorial border upheld by the user [15].

On the other hand we might interpret the raising of the hand in a “Stop” gesture as an indication to the robot that this is an advantageous position for the task at hand, i.e., the “Stop” gesture is displayed as part of a joint goal (according to [10]).



U: ok

U: now go to telephone

R: Going to telephone ①

Figure 3: An example of how communicative actions and spatial configurations of the robot are interrelated in different modalities.

Multi-modal influence (Example 3)

In the moment that passes before the example depicted in

Figure 4 the user has acknowledged that the robot has completed the task of finding an object (by establishing a common reference to the object located in front of the robot). Then the user stoops and looks into the camera of the robot, while uttering the command: “now go to the telephone”. Then, when the robot confirms the request by saying “Going to telephone”, the user changes into an upright position ① (Figure 3).

Eye contact is maintained during the whole sequence. Our analysis of this is that the user is attempting to require (visual) attention on the part of the robot. We suggest that the moving camera of the robot provides a biomimetic display that makes the user assume a transactional space located in front of the robot. In terms of Body torque, the stooping is a temporary disconfiguration of a posture, and the change back to the more neutral posture in ① (Figure 3) is a return to what Schegloff calls a home position [15]. In our understanding, the torque, i.e., the stooping posture and the user’s attempt to require contact can be contributed to the spatial influence of the transactional space [12] and the displayed “eye” gaze of the camera.

A SPATIAL PROMPTING STRATEGY

In the analysis of the scenario we have found examples that suggest that the robot platform may influence the spatial behavior, i.e., posture, gaze direction and gesture displays (e.g., stop gesture). Typically a (multimodal) natural language user interface that is used for human-robot interaction is concerned with the aspects of spatiality that are encoded in language, such as referencing using spatial relations (e.g. “behind”, “beside”, “in front of X” etc) and deictic gestures [e.g. 3].

A system that encompassed a model for spatial influence could provide *spatial prompts* aimed to influence the spatial positioning of the user, for instance, to ensure an optimal configuration for further communicative behavior. An example of such a design is depicted in Figure 4. In terms of a dialogue system design, we can frame the display of Stop gesture in the second example (Figure 2) as a system

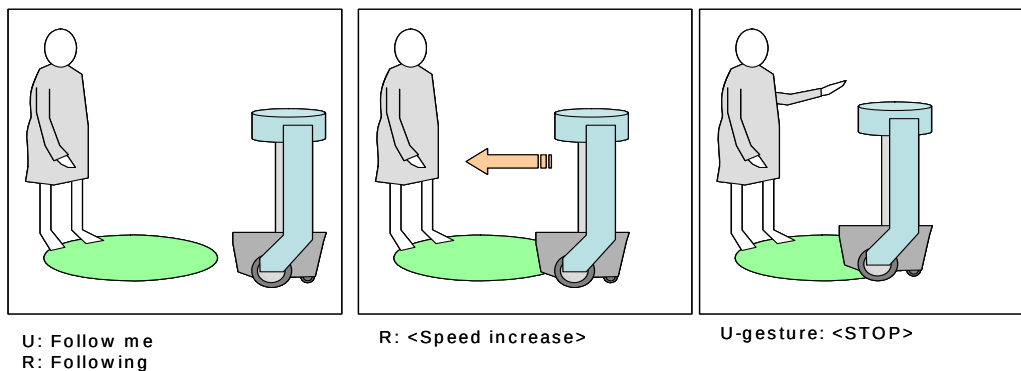


Figure 4: In the first image of the sequence, the robot is commanded to follow the user. As the user stops and the robot approaches the transactional space the robot increases its speed for a short moment. The speed increase is a spatial prompt intended to trigger a response by the user. A safety feature would stop the robot if no action is detected by the user.

goal, i.e., to reach a state in interaction where a Stop gesture has been displayed to provide an end-point in the robot's approach to a position. This requires an internal representation that considers the movement of the robot platform (and the user) together with the task state. In the follow episode in Example 2 the overall task is Follow, but as the user sits down the robot can be said to Approach the user. This could be taken into account in the system, for instance by *increasing* the speed of the robot, slightly in order to prompt the display of a gesture that shows the robot where to stop. It is obvious that a robot behavior like this raises some concerns in terms of acceptability [4], but the point of this example is merely to illustrate that spatial prompting is a possible strategy to actively influence the behavior of the users. We could also imagine an example when the robot prompts the user to tell when the robot is close enough, e.g., by saying: "say stop" while slowly approaching.

Examples like these show how we can turn empirical observations into design proposals. In terms of Body moves [9], the increase in speed in the example in Figure 4, would put Focus on the transactional space and an obligation on the user to react and specify where the robot should be positioned – before the safety feature of the robot stops it at a default distance. Then a prediction could be made, e.g., based on corpus data, so that the robot could provide a contribution that is relevant to the predicted task, e.g., "Show me an object", instead of "Stopped following" (as it is obvious to the user).

CONCLUSIONS

We have discussed a set of examples from our corpus of human-robot interactions arguing that verbal and non-verbal behavior of the robot actively influence users' spatial configuration during interaction. We also provide a scenario where a robot could utilize a spatial prompting strategy. In the future we aim to identify ways of spatially influence users by further analyze corpus data and validate the design proposal by implementing spatial prompting strategies to be tested on a robot platform.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A., and Chatila, R. (2005). Task planning for human-robot interaction. In Proceedings of the 2005 Joint Conference on Smart Objects and Ambient intelligence: innovative Context-Aware Services: Usages and Technologies Grenoble, France.
2. Pacchierotti, E., Christensen, H., and Jensfelt, P., (2005). Embodied social interaction in hallway settings: a user study. In IEEE Workshop on Robot and Human Interactive Communication Ro-Man2006, Nashville, USA, pp. 164-171.
3. Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M. and Brock, D. (2004). Spatial language for human-robot dialogs, IEEE Transactions on Systems, Man and Cybernetics, Part C., vol.34, no.2pp. 154- 167.
4. Walters, M.L., Dautenhahn, K., te Boekhorst, R., Koay K. L., Kaouri C., Woods, S., Nehaniv, C., Lee, D. and Werry, I. (2005) The Influence of Subjects' Personality Traits on Personal Spatial Zones in a Human-Robot Interaction Experiment. Proc. IEEE Ro-man 2005, pp. 347-352
5. Clark, H. H. & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language , 50(1), 62-81.
6. Gergle, D., Kraut, R. E., and Fussell, S. R. (2004). Action as language in a shared visual space. In *Proceedings of the 2004 ACM Conference on Computer Cooperative Work* (Chicago, Illinois, USA, November 06 - 10, 2004). CSCW '04. ACM Press, New York, NY, 487-496.
7. Green, A., Hüttenrauch, H., Severinson Eklundh, K. (2004). Applying the Wizard-of-Oz Framework to Cooperative Service Discovery and Configuration In Proceedings of IEEE Ro-Man, September 20-22, 2004, Kurashiki, Japan.
8. Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & Biobehavioral Reviews, 24 (6), 581–604.
9. Gill, S.P., Kawamori, M., Katagiri, Y., Shimojima, A. (2000) The Role of Body Moves in Dialogue. In RASK, April 2000, pp. 89-114.
10. Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.
11. Allwood, J., Nivre, J., and Ahlsén E. (1992). On the Semantics and Pragmatics of Linguistic Feedback, Journal of Semantics 1992 9(1):1-26.
12. Kendon, Adam. (1990). Conducting interaction – Patterns of behavior in focused encounters. Studies in interactional sociolinguistics. Cambridge, NY, USA: Press syndicate of the University of Cambridge.
13. Smith, R., Self, M, and Cheeseman P (1986) Estimating uncertain spatial relationships in robotics. UAI 1986: 435-461
14. Hall, E.T. (1966). The Hidden Dimension: Man's Use of Space in Public and Private. The Bodley Head Ltd, London, UK.
15. Sack, R.D. 1(1986). Human Territoriality: Its Theory and History. Cambridge: Cambridge University Press.
16. Schegloff, E.A., (1998). Body Torque. Social Research, 65:3, 535-596.

Measuring Up as an Intelligent Robot – On the Use of High-Fidelity Simulations for Human-Robot Interaction Research

Anders Green Elin Anna Topp Helge Hüttenrauch

School of Computer Science and Communication
KTH Royal Institute of Technology
100 44 Stockholm, Sweden Tel: +46(0)8 790 91 56
E-mail: {green, topp, hehu}@csc.kth.se

Abstract—In this article we describe and discuss how Hi-Fi simulation methods, or Wizard-of-Oz [25], can be employed in a user-centered approach to develop natural language user interfaces for robots with cognitive capabilities. By actively shaping the system both in terms of technology and in interface design the interest of both users and designers are involved in the process as stake-holders. We have analyzed two different simulation studies performed sequentially with different foci, but within the same overall scenario. The data obtained from Hi-Fi simulation studies is to a large extent qualitative, but data can also be used as a resource to be used for component development, e.g. as training data for speech recognizers. The collected data may also be used for quantitative evaluation depending on the experiment design.

I. INTRODUCTION

The creation of an interface for an autonomous robot is to some extent a very different undertaking than developing a graphical user interface for software application for the standard desktop computer.

In the European Cogniron project [1] robots with cognitive capabilities are investigated from different perspectives. The overall aim of the project is not to produce a single robot, instead several scenarios that explore and demonstrate different types of capabilities are investigated.

In this article we describe and discuss how Hi-Fi simulation methods, or Wizard-of-Oz [25], can be used in the research on human-robot interaction, where the specific aim is to explore and investigate interaction models for robots with cognitive capabilities. We have chosen to work with Hi-Fi simulations because we are dealing with technology for which users have very vague ideas, i.e. there are no cognitive service robots available on the market. Instead users' conceptions of service robots with cognitive capabilities are formed from how robots are presented in popular culture. Not even designers, i.e. roboticists, interface designers and usability experts etc, can fully predict in what manner the *robotic artifact* they are creating will be used and accepted. Maulsby et al [27] noted that “playing the role of an agent” enabled the designer to get an in-depth understanding of what type of actions the system needed to provide. This is also appears to be the case for human-robot interaction, thus in order to understand how to design the interactive behavior of a cognitive companion robot we need understand the what

this behavior should be like, i.e., formulated as a research challenge:

- *What are the requirements for the interactive behavior of a cognitive companion?*

This research challenge is not as straight-forward as it may seem. First of all the term “cognitive companion” is not well defined, but if we use the scenarios defined in the Cogniron project as a definition we end up with a system that should have the ability to use vision and audio to perceive and understand its environment, to engage in a continuous process of acquiring new knowledge and skills to learn tasks from users by interacting with them. The robot also needs to exhibit socially acceptable behaviors [1].



Fig. 1. The robot used in the two user studies in a scene from the first scenario.

The rest of this article is outlined as follows, first we describe the Wizard-of-Oz methodology and how this can be used in user-centered approaches to design of human-robot interfaces. Then we describe, compare and discuss two different studies that explore the theme of the “Home tour” scenario.

A. Wizard-of-Oz methodology

High-fidelity simulation, or Wizard-of-Oz simulation, is a methodology used for simulation of high-level functions in an interactive system. The general idea is to simulate those parts of the system that require most effort in terms of development (like a natural language understanding module) or to assess the suitability of the chosen metaphor. One of the most common uses of Wizard-of-Oz is employed for finding out how users treat a system that uses natural language as an interface.

Wizard-of-Oz simulation methodology in its classical form, i.e., where one user interacts with one (desktop) computer in a lab environment, has been used since the 1970s. The term itself was first used by Kelley [25] two decades ago. Malhotra [26] used the method for simulation of natural language expert systems. The method has also been used for simulation of database question answering systems [7], [27] and uni-modal speech interfaces [3], but also for development of multi-modal interfaces [29], [30], [32].

The starting point for a Wizard-of-Oz study involves the construction of a prototype where some features of the system are for real and where some functions are simulated by one or more operators who control the system's actions and responses.

A classical setup is to put a user in front of a desktop computer in one room and an operator, a wizard, in another room. The user is given a scenario by a test leader; a set of tasks to solve using the novel system and the interactions between the user and the system are recorded. Since the user often is unacquainted with systems of that particular kind, e.g. speech interfaces, or the task, the characteristics of the setup are that of a kind of a role-play, where the user tries to engage and act within the given scenario. Once the experiment is started the user is allowed to interact with the system in the same way as if the system was for real.

The wizard acts as the system's high-level reasoning component responding to the users actions. During the experiment the test leader may intervene if the user gets into trouble related to the use of the (simulated) system. After the user has completed the scenario the test leader normally performs some kind of post session interview. At the end of this interview the test leader briefs the user that the system was in fact simulated and asks permission to use the data that has been collected during the experiment for research purposes. If the user does not give his/her permission the recordings are noted as unanswered and the data media is erased. If there is some kind of reward to be given to the user this will be handed over irrespectively of a negative or positive reply to the request to use data.

B. User-centered design process

Hi-fi simulation is to a large extent a method that provides qualitative data on human-robot interaction. Hence, we see it as primarily important in the very early stages of the design process. During the early phases of the system development methods that can be used to communicate the intended use

of the product, e.g. mock-ups [9], [14] and scenarios with synthetic dialogues [17] are of great interest in order to inform design. When more mature prototypes have been implemented standard methods for usability evaluation [22] can be employed in order to ensure that the desired usability goals are met [8].

By now, the interaction model for a graphical user interface has become a de facto standard, and is thus known to virtually all users. The interaction model for a robot, however, is presented to the user as a *tabula rasa*, both in terms of lack of knowledge to predict the behavior of interface components and the type of tasks that the robot can perform. Normally a product that is released to the market has undergone several iterations of design and redesign. This process may be of lesser or greater complexity and the methods and work practices of may differ. However, given the fact that we in a sense are developing a "product", we may still employ methods from user-centered software engineering.

By employing a user-centered approach we implicitly accept that we are not only observing The process of developing the user interface, e.g., by taking on the role of evaluating system and human performance along the lines of [35]. Instead we are actively shaping the system both in terms of technology [36] and in interface design [14], [20], hence both users and designers have a role as stake-holders in this process. Thus we need to identify values that we aim to support with our systems [12], as well as to understand that we as researchers are part of the process [4] and not merely evaluating a system that has been created by others.

C. Hi-Fi Simulation studies of Human-Robot Interaction

In the Human-Robot Interaction domain there are some compelling reasons for performing simulations rather extracting data from human-human communication.

In many of the envisioned scenarios human-robot communication is focused on communication about topics that are already known to the participants during conversation, making it unnecessary to address these explicitly using speech during the execution of a task. First of all there is the matter of the task that the robot is performing. The tasks that we address the robot with are either simple or focused on domains where explicit verbal instructions are rare or only sparsely used by humans, but seem very important for robots. For instance, explicit negotiation of easily understood tasks (e.g. fetching an object from a known location) [17], and communication concerning detection of humans and description of routes [13]. Secondly, even in scenarios where robots are intended to replace people, in dirty, dangerous or distant environments, the work allocation between actual workers and robots may differ (see [6] for an example of extreme use). This means that the conversation between field workers, like divers or rescue workers would probably be of little direct use to when developing modules for spoken human-robot communication.

We may also perform simulation studies in order to investigate hypotheses concerning general aspects of human-

robot interaction, such as social behaviors, e.g., studies of spatial positioning [40], [41] and collaboration [21].

As a method Hi-fi simulation supports a wide variety of activities where designers and users increase their understanding of a future system. According to [15] these activities can be categorized along the dimensions try-out, training and testing:

- *Try out and conceptualize design ideas*, i.e., to rapidly set-up and perform formative studies in order to conceptualize and try out new ideas for the interface that is being developed.
- *Training and education*, i.e., to train natural language components such as speech recognizers. It also provides an opportunity to educate interface designers by letting them to take on the role of the system [27], and thereby increase their understanding of the task.
- *Testing, evaluation and study of use*, using standard usability methods for the type of interface that is being evaluated, for instance using the “Common metrics” approach according to [35] or evaluation methods that are especially targeted for natural language user interfaces, e.g., PARADISE [39]. Last but not least, we include the study of systems in realistic usage contexts, i.e., observations that can be analyzed to increase our understanding of human-robot interaction in general.

II. WIZARD SIMULATION STRATEGIES

When employing Hi-Fi simulation techniques, such as Wizard-of-Oz, we need to establish what kind of data we are aiming to collect and how it should be analyzed. When we are simulating an autonomous robot with the ability to communicative in natural language we have identified a set of dimensions, in terms of behaviors and appearance, along which the system can vary:

- Degree of system realism: Are we going to simulate a “realistic” system? We must decide to what extent we are going to simulate system behavior so that they appear as realistic to the user. For instance, if we are simulating navigation capability we need to assure that the wizard does not “equip” the system with a navigation system that is able to plan ahead, seemingly inferring the intent of the user. Hence it is important to provide a set of constraints that bring some realism into the situation of use. This is what Maulsby [27] refers to being “true to the algorithm”.
- Degree of exploratory freedom: Are we going to simulate according to an “algorithm”? Decisions that are made along this dimension concern whether or not we should allow a completely free and explorative interaction style, or if we should restrict the task
- Behavioral mimicry: how “natural-like” should the system be in terms of appearance and interactive behaviors, i.e., to what extent should the system imitate “nature”, in terms of appearance (e.g., using close resemblance with nature [31]) or interactive capability, e.g., allowing for a conversational style of interaction that is close to human capability.

A. Validity of Wizard-of-Oz simulation

Sometimes the Wizard-of-Oz method is criticized for not providing necessary data for evaluating practical dialogue. Wizard-of-Oz methods typically fail to involve users that bring real tasks to the system [23]. This should be seen in contrast to what is generally believed, namely that the Wizard-of-Oz method is an open-ended method for collection of data on user behavior. Allwood and Haglund [2] have noted that the wizard operator acting a scenario is involved in roles on different levels. Thus the researcher role involves acting as a system (wizard operator) and during sessions the wizard can take on different communicative roles like the sender role, the receiver role etc. Bell [5] notes that in a task scenario, like a travel agency dialogue, the wizard not only acts in a system role but in the role of a travel agent. In reality the behavior of people acting in a real situation may be quite different from what people do in a simulated scenario even if the user believes that she is interacting with a real system.

When simulating a multi-modal system with many different ways for the user of providing input the cognitive load of the operators increases. There is sometimes a mismatch between what needs to be simulated and what suits the cognitive and perceptual abilities of the wizards. This phenomenon was been noted by Fitts [10] who proposed a list describing what people do better than machines and what machines do better than people¹. In a simulation scenario, the problem of function allocation is twofold. First of all we need to consider the characteristics of the system we are simulating and secondly we need to think about the function allocation of the system we are using for the simulation. Following Sheridan [33] we see may see Fitts’ List as a set of accepted statements making up the foundation for how to reason when designing function allocation. According to Fitts people are better than machines at detecting small changes in the environment, perceiving patterns; improvising, memorizing, making judgments and reasoning inductively. Machines are better at responding quickly to signals, applying great force, storing and erasing information and reasoning deductively [33]. It is therefore important that we consider the strengths and weaknesses of machines and humans when designing and planning Wizard-of-Oz setups.

Ethical considerations

The use of the Wizard-of-Oz method provides a possible ethical dilemma for the researcher, since in the classical setup, the user is led to believe that the system is a real machine. When the truth is revealed afterwards, the test-leader normally asks for the permission to use the data. There are a number of studies that use deception; in fact it seems that the use of the term Wizard-of-Oz is always connected to the practice of deceiving the user. Fraser and Gilbert [11] have argued against deceiving users on ethical grounds. This ethical dilemma might be solved by telling the user that the system is simulated, making the setup part of

¹The so-called MABA-MABA list.

the role-play. The validity of the results from a study that is performed in this overt fashion may be questioned but this is dependent on the purpose with the study. Dahlbäck et al [7] argue that for some aspects of dialogue, like the type and frequency of anaphoric references, it is important that users are deceived rather than engaging in a role-play type of scenario. We should note that the work by Dahlbäck [7] and others in the late eighties was mostly done with systems that used text input. For spoken scenarios role-playing has been considered useful in multi-user scenarios [24] and elicitation of error handling strategies [34]. Human natural language performance is automated to a great extent and therefore we may assume that phenomena on the low-level like syntax, and vocabulary used probably are little affected by the fact that the user knows that the system is simulated or not.

However, using the Wizard-of-Oz method in the classical, deceptive manner, ensures the illusion of speaking with a computer. In a non-deceptive Wizard-of-Oz study, this illusion might be broken, something which may have an adverse effect on the result. However, it seems strange to use made up scenarios and role-play while maintaining the standpoint that we should deceive the user to the extent that the system is in fact real. In our view is that it is still an open question if deception is really necessary?

III. STUDYING THE “HOME TOUR SCENARIO”

The two studies that we are discussing in this paper both stem from work that is performed in the context of the European project Cogniron [1] where a set of scenarios has been defined. In the scenario we have worked with primarily, the user gives the robot a “tour” of the environment:

In this experiment, a robot discovers a home-like environment and builds up an understanding of it and of artifacts in it as taught by humans. This process is open-ended, i.e., it has no completion: the robot continues to learn as it faces new situations. A human shows and names specific locations, objects and artifacts, to the robot. The robot can engage in a dialogue in case of missing or ambiguous information.

This scenario can be characterized as kind of Co-operative Service Discovery and Configuration, stressing the way the user and robot is intended to engage in a joint effort to inform each other of relevant knowledge about the environment. The central themes that the scenario aims to address concerns two main types of information to be jointly discovered by the user and the robot:

- (i) the artifacts present in the environment (e.g. objects and locations) and,
- (ii) the actions that the robot can perform related to these artifacts.

The two studies that we discuss in this paper explore two key concepts. We have already introduced *Cooperative-Service Discovery and Configuration* which can be viewed as an abstraction of the scenario described above. In the second study the notion of Human-Augmented Mapping (HAM,

[36]) was evaluated both in terms of interaction and from a technical point of view. In the following we will compare the set-up and the result of these studies. An overview is given in the Tables I and II.

The general purpose of the first study was to evaluate an extended dialogue model based on [19], [38] in order to provide a rich set of data for further analysis of human-robot communication. The first study was primarily aimed at exploration and not formal hypothesis testing. We did however put down a list of things that we believed that users would do in different situations during the study in order figure out what the wizards should do with the aim to provide a consistent behavior throughout the session.

In the second study, which can be characterized as a pilot study, the general aim was slightly different. We still have an explorative purpose based on the assumption that individuals have different preferences and strategies for how to present their everyday environment. The second study can also be seen as a technical test of a way of represent the environment. Using the notion of MacNamara [28]) we assume that individual strategies are based upon a common psychological representation the meaning and conventionalized use of the terminology that describe users environment. Thus, during the sessions, we use the communicated concepts to annotate the environment representation (i.e., nodes with the types location or region). This means that the role of the wizard is to decide whether the communicated concepts, e.g. phrases like “kitchen” and “my office”, can be categorized as locations and regions according to the environment representation. Study two thus aims to answer to questions:

- What strategies are used by individual users to *present* the environment?
- How can the information given by users be incorporated in a representation that can be used as a shared representation.

Both studies can be characterized as Hi-Fi simulations, the first study was performed as pure wizard study, i.e, users were not informed that the system was controlled by a wizard operators. The wizard operators were not completely hidden, instead the users were told that they were “technical support staff” that “monitored the experiment”.

When there are two wizard operators it is necessary to carefully consider the task allocation between wizards and to support their collaboration. In the first study there were two wizards, one controlling the dialogue (communicator) and one controlling the movements of the robot (navigator). The wizards stood close together during the session to support contextual awareness, meaning that they easily could get a glimpse of what the other wizard was doing while concentrating on the task. The setup can be seen in Figure 2.

In the second study the users were informed that the speech system was controlled by the accompanying operator. Whereas the first study was performed with 22 invited users (undergraduates from our campus) and thus did not know the environment, the 5 users in the second study were all staff, familiar with the lab environment. As familiarity with the environment was a requirement depending on the research

questions posed in the second study lab staff was chosen. In future studies we aim to move the robot to an environment outside the lab so that users that have little experience with robotics can be studied.

The first study was performed in a single room in our lab (Figure 2), furnished like a living room to provide a home-like character. The second study was performed in a significantly larger environment, i.e., a whole floor of our lab comprising about 20 rooms (offices and common areas like kitchen, meeting room and printer area).

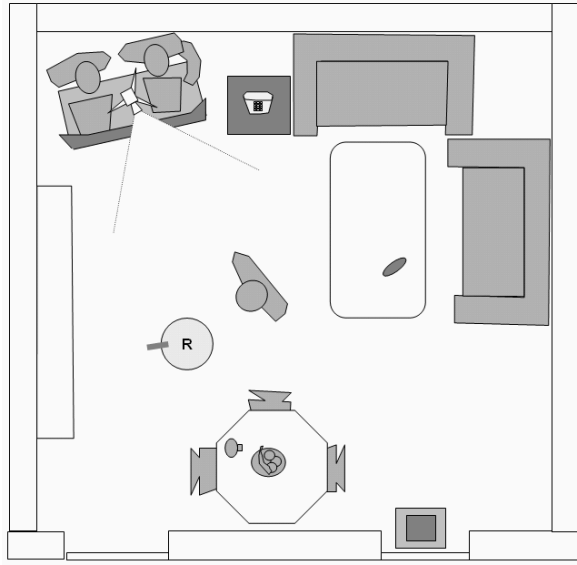


Fig. 2. Map showing the single room used in Study 1

a) Robot appearance, behavior and instructions to users: In both studies we used an ActivMedia PeopleBot equipped with an on-board pan-tilt-camera (see Figure 1. In the first study the robot also was equipped with visible stereo microphones.

In the first study, in terms of behavior the movements (navigation and camera movements) of the robot was controlled by one wizard and the speech output was controlled by another. Since we were aware that a real robot navigation component would make the robot behave in a specific way we tried to mimic this behavior as far as it was possible. In the second study a real follow behavior was used to solve all tasks, except for error recovery, which was performed overtly as tele-operation.

In terms of interactive behavior the first study was more focused on providing a coherent and consistent dialogue behavior, i.e., the dialogue wizard tried to follow an “algorithm” using the terminology by [27]. Hence the subjects received instructions with clear examples of how they could use the system. Our goal was not to test the system “in the wild” without informing the users about how the system was intended to work. Instead we very much instructed the users so that they would be able to use the system. Hence, the instructions were intentionally clear with exact phrasing. The navigation wizard, who also acted as the test leader, also

showed the user how the robot worked by giving a small demonstration.

The robot had a dialogue model that accommodated for the actions:

- Greeting, e.g., phrases like “hello” and “my name is john”.
- Follow, e.g., phrases like “follow me”
- Demonstration, e.g., referential phrases and gestures like “this is an orange” + pointing gesture.
- Validation, e.g., phrases like “find the orange!”

For the second study the robot used two different ways of giving feedback on users’ input, dependent if the user provided information for labeling a *location* or a *region*. Thus if a location, e.g., “coffee machine” or “door to” a *region*, was presented by the user the robot would not move and immediately state that it had stored the location. If a region, e.g., “kitchen”, “my office”, was presented, the robot would state that it needed to “look around” and make a full 360 degree turn, before confirming that a *region* had been stored. The role of the wizard in was to select which behavior to use based on the wizards understanding if whether the concept used by the user could be interpreted as referencing a *location* or a *region*.

Users were told to use commands, i.e., follow me, *go to* <target>, *stop*, *turn* <left/right>, to control the robot and referential expressions like *this is* <term> together with deictic gestures. No terminology that could be directly related to the concepts tested was mentioned to the users, e.g., *location*, *region*, *place* etc. Instead phrases like “everything, that you think the robot needs to know” was used to guide the user. Since a real follow behavior was used special instructions was needed in order that the system would be able to track and follow the user (i.e., the user had to move about 1 meter to make the robot start following).

b) Data collection setup : In the first study we could collect data in a fairly controlled manner using a video camera placed in the location of the users. We also had web-cameras placed in each corner of the room. Mounted on-board the robot an audio recorder captured high quality stereo sound. Data from the laser range finder was also stored.

During the second study we used a hand-held camera and on-board camera to capture the “scene” of the user and the robot acting together in the environment. Since the environment comprised over twenty offices and common areas (e.g. kitchen, printer area) we could not cover the whole area with web cams or sensors.

In both studies we administered a questionnaire and interviewed the users after the end of the trial. In the first study we asked the users about who was controlling the robot before informing them about that they had been involved in a wizard study and asked again about their willingness to participate given this new information. No user had any objections once we had revealed the truth and explained the reasons for our conduct. In the second study the subjects were informed that the system was partially simulated. The post-session interviews in the second experiment also included a sketch

Study type	Study 1: Single room	Study 2: Multiple rooms
Purpose	Data collection and exploration of interaction strategies related to a specific dialogue model [19], [38]	Investigate interaction strategies used by individual users to <i>present</i> the environment
Type, size of study	Explorative, 22 subjects (20+20 min)	Pilot-study with 5 subjects (20+25 min)
Key concept	Co-operative Service Discovery and Configuration	Human-Augmented Mapping (cf. [36])
Type of subjects	Undergraduate students	Staff that knew their environment
Environment	Single “living room”	20 rooms + kitchen

TABLE I
GENERAL PURPOSE AND SETUP OF USER STUDIES 1 AND 2

Interaction	Study 1: Evaluation of Dialogue model	Study 2: Presentation strategy
Robot behavior and appearance	Dialogue model accomodating tasks: greet, follow, show (object, location), validate, close	Two behaviors: show (acknowledge) and show (look-around)
Instructions	User were provided with a written instruction with examples. Testleader gave a demonstration.	Users were told to use a small set of commands and gestures
Wizard-status	Users were informed about the role of robot operator	Pure wizard – user were told that wizard was an operator
Number and role of wizards	Two wizards (navigator and communicator) behind a screen (cf. [15])	One wizard accompanied user/robot
Data collection setup	• Video camera (overview). Onboard stereo audio (robot perspective). • Network web cameras in each corner. • Questionnaire • Post-trial interview	• Hand-held camera (overview) / onboard camera (robot perspective) • Questionnaire • Post-trial interview (scripted) • Sketch session
Type of data	• Single audio/video track (Mini DV) • Web-cam images (~ 1 fps) • Data from laser range finder • Stereo audio (22 KHz) • Notes from interviews • Questionnaire data	• Two video tracks with audio (Mini DV) • Notes from interviews • Data from laser range finder • Environment representation • Questionnaire data • Video from sketch session (Mini DV)

TABLE II
INTERACTION DIMENSIONS OF USER STUDIES 1 AND 2

session where the user was instructed to draw a map on a white-board and explain how they had instructed the robot.

A. Findings from the studies

Several analyses have been performed using the data collected. An analysis of the communicative acts have been made and reported in [16]. We have also studied the role of posture and positioning [20] and active spatial influence on the part of the robot [18].

We performed an analysis of the miscommunication using the video recordings from the first 12 user sessions. These were transcribed and synchronized on the utterance level. We then printed out all dialogues and analyzed them by marking utterances that could be interpreted as symptoms of miscommunication.

We found miscommunication that could be attributed to the users’ erroneous inferences about the system’s capability. For instance, 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.

We also 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. 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. Overlapping speech was also a problem in the

corpus, i.e., the robot should have stopped the synthesized speech when the robot started to speak, something which lead to misunderstandings.

Another problem that can be described as an effect of the modality is miscommunication related to 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. 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 on-board camera, there is no way of indicating precisely which object has been detected.

Results from the second study have been reported in terms of a technical evaluation of the environment representation [37]. The strategies employed by the users were very different and no convergence in behavior could be established since there were only five subjects. The observed behaviors still need to be taken into account when constructing future systems. Some of the phenomena are provided below to give a sense of what went on during the sessions:

- Users' personal view of the environment affects the way it is introduced to the robot, e.g., a person not drinking coffee might not mention the coffee maker.
- Persons were pointed out to the robot as offices (with occupants in them) were passed. Bystanders that passed the robot were also pointed out by one user.
- Locations pointed out reflected an action oriented view of the environment, i.e., a coffee machine is mentioned as a place where the robot should get coffee, but the kitchen is not mentioned.
- Doorways were pointed out by subjects with knowledge about robot navigation
- After giving the follow-command, the users stood still and waited for the robot to move, something which was not anticipated. Hence it needs to be equipped with an active way of prompting the user to move (i.e., perform an act of *spatial prompting* [18]).

IV. DISCUSSION

When looking at the two different studies in a bird's eye perspective, the generic aspects of both studies can be contributed the context of use provided by the overall scenario, the Home Tour:

- The use of natural language using speech and gestures
- The initiative during the interaction lies with the user
- The behavior of the robot influences the user both in terms of communicative behavior and spatial adaptation.

In some respect the different foci of the studies affect the way studies have been setup. When mobility of both users and systems become a topic for investigation, the complexity of setting up the scenario increases, something which have an effect the way data can be collected.

Clearly the wizard-of-oz framework can be employed successfully for simulating a full fledged interactive robot system. There are several dimensions that have been added

since the early work on screen-based applications [7], [25]–[27]. The dimension of collaborative behavior has always been in focus of wizard-of-oz studies for natural language user interfaces. For human-robot interaction the situated use context is in focus, i.e., the human and robot share the same environment and partially share the same perceptual context.

Since a robot system allow several degrees of freedom the manner and ability for the wizards to control the robot system almost becomes research topic in itself. It is clear that to some extent we need to assure that we do not spend more resources on building a simulation environment than on building the real system. The use of a real robots also makes safety issues come in focus, hence we need to assure that we can maintain safety during the sessions. We also need to check that the more formal aspects of user studies, i.e., make certain that legal and ethical issues are compatible with the type of studies we will perform (see Walters et al [40] for a specific example from the UK).

V. CONCLUSIONS

In this article we have described the way in which Hi-Fi simulation studies can be used in the process of developing interactive service robots. When carefully designed, simulation studies will provide data about different aspects of human-robot interaction that would otherwise be unattainable until large effort had been spent on the creation of a working prototype. Thus, the data that we get from a Hi-Fi simulation study is to a large extent qualitative, but we may also collect data as a resource to be used for component development, e.g. as training data for speech recognizers. We may also use the collected data to perform quantitative evaluations using different standard methods [35].

The type of data we can get a wizard-of-oz type simulation study ranges from collection of data that can be analyzed in different ways.

- Data on language use and task strategies, especially spatial language (including gesture and speech) can be analyzed with methods from psychology, linguistics etc.
- Visualization of the whole interaction to enable the designer to conceptualize what the system could or should do in different realistic situations.
- Assessment of users' attitudes towards a future system or towards robots in general.

In the future our aim is to evaluate systems where tasks that now are handled by wizard operators gradually are replaced with real components. To succeed with this we need investigate how we can provide a system which *can* be managed by wizards, while providing the necessary exploratory freedom for users together with requirement of an appropriate degree of system realism.

VI. 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.

REFERENCES

- [1] COGNIRON, Annex 1 - Description of Work, 2003. EU Sixth Framework Program FP6-IST-002020, <http://www.cogniron.org>.
- [2] J. Allwood and B. Haglund. Communicative Activity Analysis of a Wizard of Oz Experiment. Technical report, Department of Linguistics, Göteborg University, 1992.
- [3] G. Antoniol, R. Cattoni, M. Cettolo, and M. Federico. Robust Speech Understanding for Robot Telecontrol. In *Proceedings of the 6th International Conference on Advanced robotics*, pages 205–209, Tokyo, Japan, 1993.
- [4] L. J. Bannon and S. Bædker. *Designing Interaction: Psychology at the human-computer interface*, chapter Beyond the Interface: Encountering Artifacts in Use. Cambridge University Press, New York, 1991.
- [5] L. Bell. *Linguistic Adaptations in Spoken Human-Computer Dialogues: Empirical studies of User Behavior*. PhD Thesis, KTH Royal Institute of Technology, 2003. TRITA-TMH 2003:11.
- [6] J. Casper and R. R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics Part B*, 33(3):367–385, 2003.
- [7] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of Oz studies - why and how. *Knowledge-Based Systems*, 6(4):258–256, 1993.
- [8] P. Dario, E. Guglielmelli, and C. Laschi. Humanoids and personal robots: Design and experiments. *Journal of Robotic Systems*, 18(12):673–690, December 2001.
- [9] P. Ehn and M. Kyng. *Cardboard computers: mocking-it-up or hands-on the future*, pages 169–196. Lawrence Erlbaum Associates, Inc., 1992.
- [10] P. M. Fitts. *Human engineering for an effective air navigation and traffic control system*. National Research Council Committee on Aviation Psychology, Washington, DC, 1951.
- [11] N. M. Fraser and G. N. Gilbert. Simulating Speech Systems. *Computer Speech & Language*, 5(1):81–99, 1991.
- [12] B. Friedman and P. Kahn. The human-computer interaction handbook. chapter Human values, ethics, and design. Lawrence Erlbaum Associates, Mahwah, NJ, 2003. Eds: J. A. Jacko and A. Sears.
- [13] J. Fry, H. Asoh, and T. Matsui. Natural Dialogue with the JJO-2 Office Robot. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2:1278–1283, 1998.
- [14] A. Green, H. Hüttenrauch, M. Norman, L. Oestreicher, and K. Severinson Eklundh. User-Centered Design for Intelligent Service Robots. In *Proceedings of 9th IEEE International Workshop on Robot and Human Interactive Communication*, Osaka, Japan, 2000.
- [15] A. Green, H. Hüttenrauch, and K. Severinson Eklundh. 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*, pages 575–580, 20–22 Sept 2004.
- [16] A. Green, H. Hüttenrauch, E. A. Topp, and K. S. Eklundh. Developing a Contextualized Multimodal Corpus for Human-Robot Interaction. In *Proceedings of the Fifth international conference on Language Resources and Evaluation LREC2006*, 2006.
- [17] 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.
- [18] A. Green, B. Wrede, K. S. Eklund, and S. Li. Integrating Miscommunication Analysis in the Natural Language Interface Design for a Service Robot. submitted to IROS2006, 2006.
- [19] A. Haasch, S. Hohanner, S. Huewel, M. Kleinhagenbrock, S. Lang, I. Tóptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer. BIRON – The Bielefeld Robot Companion. In *Proceedings of ASER 2004 - 2nd International Workshop on Advances in Service Robots*, Stuttgart, Germany, May 21 2004.
- [20] H. Hüttenrauch, A. Green, M. Norman, L. Oestreicher, and K. Eklundh Severinson. Involving Users in the Design of a Mobile Office Robot. *Systems, Man and Cybernetics, Part C: Applications and reviews*, 34(2):113–124, 2004.
- [21] H. Hüttenrauch and K. Severinson Eklundh. To Help or Not to Help a Service Robot. In *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication RO-MAN'2003*, Millbrae CA, USA, 2003. IEEE.
- [22] ISO. International Standard ISO/IEC 9126. Information technology - Software product evaluation - Quality characteristics and guidelines for their use, International Organization for Standardization. 1991.
- [23] A. Jönsson and N. Dahlbäck. Distilling dialogues - A method using natural dialogue corpora for dialogue systems development. In *Proceedings of 6th Applied Natural Language Processing Conference*, pages 44–51, 2000.
- [24] K. Kanto, M. Cheadle, B. Gambäck, P. Hansen, H. K. Kristiina Jokinen, and J. Rissanen. Multi-Session Group Scenarios for Speech Interface Design. In C. Stephanidis and J. Jacko, editors, *Human-Computer Interaction: Theory and Practice (Part II)*, volume 2, pages 676–680. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2003.
- [25] J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41, 1984.
- [26] A. Malhotra. Design criteria for a knowledge-based English language system for management: an experimental analysis. Technical Report MAC TR-146, MIT, 1975.
- [27] D. Maulsby, S. Greenberg, and R. Mander. Prototyping an Intelligent Agent through Wizard of Oz. In *INTERCHI'93*, pages 277 – 282. ACM, April 1993.
- [28] T. McNamara. Mental representations of spatial relations. *Cognitive Psychology*, 18:86–121, 1986.
- [29] S. Oviatt. User-centered Modeling for Spoken Language and Multimodal Interfaces. *IEEE Multimedia*, 3(4):26–35, 1996.
- [30] D. Perzanowski, A. Schultz, W. Adams, E. Marsh, and M. Bugajska. Building a multimodal human-robot interface. *Intelligent Systems, IEEE [see also IEEE Expert]*, 16(1):16–21, 2001.
- [31] T. Saito, T. Shibata, K. Wada, and K. Tanie. Relationship between interaction with the mental commit robot and change of stress reaction of the elderly. In *Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation*, volume 1, pages 119–124, 16–20 July 2003.
- [32] D. Salber and J. Coutaz. Applying the Wizard of Oz Technique to the Study of Multimodal Systems. In *EWHCI*, pages 219–230. 1993.
- [33] T. B. Sheridan. Function allocation: algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies*, 52(2):203–216, 2000.
- [34] G. Skantze. Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems. In *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 71–76, 2003.
- [35] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. Common Metrics for Human-Robot Interaction. In *In Proceedings of HRI2006 1st annual conference on Human-Robot Interaction*, Salt Lake City, UT, USA, March 2–3 2006. ACM.
- [36] E. A. Topp and H. I. Christensen. Tracking for Following and Passing Persons. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005)*, Edmonton, Alberta, August 2005.
- [37] E. A. Topp, H. Hüttenrauch, and H. I. C. K. Severinson Eklundh. Acquiring a shared environment representation. In *Proceedings of the 1st ACM Human Robot Interaction Conference, HRI2006*, Salt Lake City, Utah, USA, March 2006. (extended abstract).
- [38] I. Tóptsis, S. Li, B. Wrede, and G. A. Fink. A Multi-modal Dialog System for a Mobile Robot. In *INTERSPEECH: 8th International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 273–276, Jeju, Korea, 2004.
- [39] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 271–280, Madrid, Spain, July 1997.
- [40] M. Walters, S. Woods, K. Koay, and K. Dautenhahn. Practical and methodological challenges in designing and conducting interaction studies with human subjects. In *Proceeding of AISB Symposium on Robot Companions Hard Problems and Open Challenges in Human-Robot Interaction*, pages 110–120, Hertfordshire UK, April 2005 2005.
- [41] M. L. Walters, K. Dautenhahn, K. L. Koay, C. Kaouri, R. te Boekhorst, C. L. Nehaniv, I. Werry, and D. Lee. Close encounters: Spatial distances between people and a robot of mechanistic appearance. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots (Humanoids2005)*, pages 450–455, Tsukuba, Japan, December 5–7 2005.

Developing a Contextualized Multimodal Corpus for Human-Robot Interaction

Anders Green, Helge Hüttenrauch, Elin Anna Topp, Kerstin Severinson Eklundh

Royal Institute of Technology
School of Computer Science and Communication
100 44 Stockholm
Sweden
{green, hehu, topp, kse}@csc.kth.se

Abstract

This paper describes the development process of a contextualized corpus for research on Human-Robot Communication. The data have been collected in two Wizard-of-Oz user studies performed with 22 and 5 users respectively in a scenario that is called the Home Tour. In this scenario the users show the environment (a single room, or a whole floor) to a robot using a combination of speech and gestures. The corpus has been transcribed and annotated with respect to gestures and conversational acts, thus forming a *core annotation*. We have also annotated or linked other types of data, e.g., laser range finder readings, positioning analysis, questionnaire data and task descriptions, that form the annotated *context* of the scenario. By providing a rich set of different annotated data, the corpus is thus an important resource both for research on natural language speech interfaces for robots and for research on human-robot communication in general.

1. Introduction

The purpose of this paper is to describe a corpus which is used in the research on cognitive robots in the European project Cogniron¹. We will also describe the development process and challenges involved when collecting and annotating the corpus, and the way we are able to contextualize the different types of data. One important aim with the corpus is to support the development of natural language user interfaces for a robot with cognitive capabilities.

We are striving to collect data from many different sources in order to be able to provide a rich context for the modalities that are used for interaction in order to provide means of analyzing the data from different perspectives. Thus we have annotated communicative actions: speech and gesture and other actions related to the task and spatiality: data on positioning of users, objects and locations. This should be seen in contrast to the corpus developed by Maas and Wrede (2006) which focuses on capturing higher dialogue structures (i.e., topics) that emerge during human-robot interaction. Both efforts are in the long run aimed to enable users to train the robot to perform a wide range of tasks that are not preprogrammed – using a multimodal style of interaction. 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 (Green et al., 2006) and to analyze users’ positioning (Hüttenrauch et al., 2006) and task strategies.

1.1. Related research

There are initiatives to collect corpora for multimodal interfaces (Knudsen et al., 2001; Schiel et al., 2002) but few that are targeted for robotics (Bugmann et al., 2001; Bugmann et al., 2004; Wolf and 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

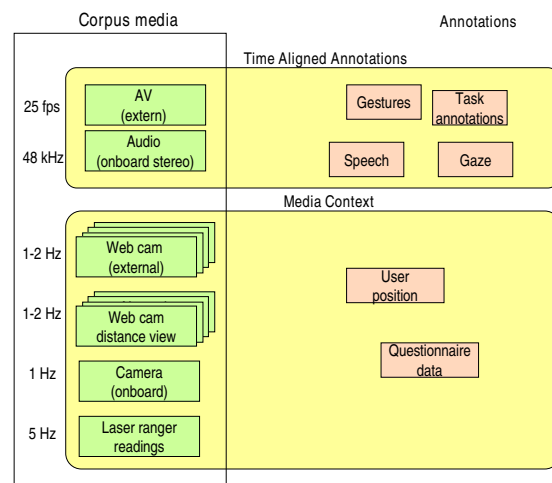


Figure 1: An overview of the corpus

observations of user behavior, e.g., gaze behavior, to evaluate human engagement in interaction (Sidner et al., 2004).

2. General scenario for data collection

The current version of the corpus contains data collected from two user studies (Green et al., 2004; Topp et al., 2006) that have been set up in order to explore user behavior in a scenario that can be characterized as a “Home Tour”. 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.

This scenario can be characterized as kind of Co-operative Service Discovery and Configuration (Green et al., 2004), 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 *discover* what the robot can do, and is able to *configure* it by actively providing information about the environment. In the studies described in this paper the user can specify

¹www.cogniron.org

names of objects and locations. But in a longer perspective, the user should also be able to interactively specify the tasks the robot can perform related to these locations and objects. In the home tour scenario the user is to guide the robot in an environment containing different objects that could potentially be recognized by the system. Thus the main task for the user is to *introduce herself* to the robot and to *show it objects and locations*. To move the robot around a follow-behavior is used to position the robot during the task. To collect data and explore the character of such an interaction we designed and set up a tele-operated robot system that could be used to perform a Wizard-of-Oz simulation, a technology that has been described in more depth by, e.g., Dahlbäck et al. (1993).

The robot we used for the trials was a ActivMedia PeopleBot (Figure 2). The pan-tilt camera mounted on the robot was moved by the wizard during the sessions so that it appeared as looking at the things that were specified by the user.

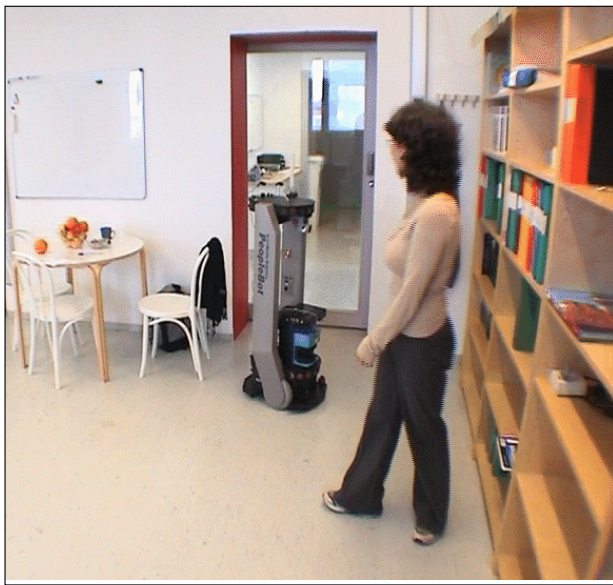


Figure 2: The modified ActivMedia PeopleBot used for the corpus collection and a user engaged in interaction.

2.1. User study 1: Single room – constrained dialogue model

For the first user study we used a room in our robot lab (see Figure 2 and 3) which had been equipped with a set of furniture: a sofa, a dinner table with some chairs, a tv set, book shelf, objects on the table (fruit bowl, mobile phone) etc.

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 (9 female, 13 male, ~24 years old), but since the aim of the study is primarily explorative we have accepted this circumstance.

2.1.1. Instructions to users

When a user arrived, 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”. During the trial there were three researchers present; one acting as test leader/navigator; one acting as communicator; and one acting as observer. During the setup the observer was positioned in one of the sofas taking notes.

After the introduction the subject signed an agreement giving consent to storing of personal information. The instruction to the user was first of all provided as demonstration, where the test leader addressed the robot, made it follow (i.e. by saying “follow me”); showing it an object by pointing and saying “this the green book”). Then the test leader commanded the robot to get back to the starting position by saying: “go to the recharge station”. The user was also given a written instruction describing the task and principal services the robot could perform:

Task: The user was instructed to use the available dialogue capabilities to teach it objects and locations that were depicted on the back of sheet of paper they were holding.

Following: The follow behavior was described to the user by providing an explicit example of what to say, i.e., “Say ‘Follow me’ to make the robot follow you”.

Showing objects and locations: The Show task was described in an indirect way, i.e., not providing any explicit phrases to the user, with the aim to avoid priming of lexical choice:

You may use your hands to show a single object to the robot. Objects that the robot should know can be indicated if they lie on a flat surface like a coffee table. The surface need to be free from other objects – the robot will use its vision system to collect information about the objects. Say the name of the object that the robot should learn to the robot and use your hand to point where it is.

These descriptions should work as both an aid for the wizard and a constraining factor for the scenario. The underlying assumption for introducing the user to a simulation of a natural language user interface is to provide the freedom to interact in a way that seems natural to the user – without actually implementing the system for real. However, it is important to provide a set of constraints that bring some realism into the situation of use. This is what Maulsby et al. (1993) refer to being “true to the algorithm”.

2.2. User study 2: multiple rooms – less constrained dialogue

The home tour scenario described earlier is also relevant to the concept of Human Augmented Mapping (HAM) introduced by Topp and Christensen (2005) the aim of which is to provide a link between human-robot interaction and robotic mapping in a way that is compatible with human cognitive representations.

To explore this scenario we designed another user study where the environment where the user and robot interacted was larger. In this case we extended the experiment area to a whole floor of the robot laboratory. This was done in order to provide a scenario that is sufficiently complex both from a technical point of view, i.e., where data collection is

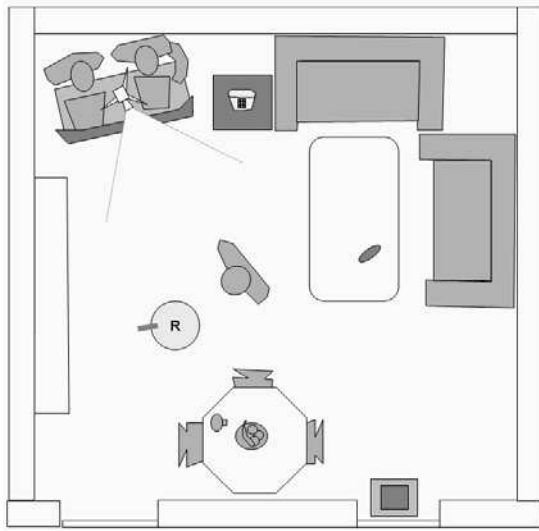


Figure 3: A map of the room used in the data collection. In one corner the position of the wizards is shown. The different objects, like the fruit bowl and the remote control, were always placed on the same initial positions before the study started.

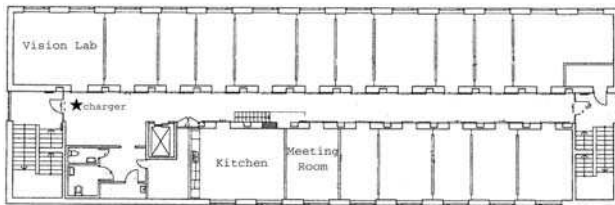


Figure 4: The floor plan of the office environment where the experiment took place. The most prominent places was the kitchen and the robot lab.

used to evaluate algorithms, and provide data on interaction between robot and user.

For this study, which is still ongoing, we have only recruited five users, where knowledge of the environment was a requirement. The users were instructed to use the follow behavior and to present the environment with respect to what locations that the users perceived as important for the robot to know.

The instructions did not include explicit directions on how to name locations to the robot, this was left for the user to find out. The interaction provided by the robot was limited to acknowledging: a) that a location had been received and b) that a follow task has been initiated ("robot is following"). The interaction was recorded using a hand-held video camera and an video camera placed onboard the moving robot, providing a "robot perspective". After the experiment the subject was interviewed by the experiment leader.

3. Corpus annotation

The recorded video material from both studies is being digitized, transcribed and annotated along several dimensions to support usability research and development of

cognitive modules. The annotations fall into two broad categories: annotation of *communicative acts* and *supportive* or *context-providing annotations*. This is dependent on how the specific type of annotation can be used to perform analyses without using other data. We should note that what is regarded as context is of course dependent on what the analysis is focusing on. However, we have noted that annotations that fall in the *core* category, i.e., time-aligned speech, communicative acts and gestures are invaluable for navigating in the material, e.g., when looking at distances (Hüttenrauch et al., 2006).

3.1. Annotation of Communicative acts

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 into Anvil XML files (Kipp, 2004) allowing the sessions to be displayed in several layers.

We are using a coding taxonomy to capture communicative acts that can be viewed as multimodal extension of the DAMSL coding schema (Allen and 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 analysis. Our approach is similar to Villaseñor et al. (2000), who proposes the extension of DAMSL with the notion of contribution as participatory communicative acts, according to (Clark and Schaefer, 1989).

Report-task: The categories Report-Task and Report Task-fail were primarily asserted to utterances where the robot provides a report concerning the task, much like a comment, e.g., "Robot is following" or "Cannot do that". Allen (Allen and Core, 1997) classifies utterances related to the task on the Information-level, using the categories Task and Task-Management. We have chosen not to annotate the Information-level in our corpus, since the style of interaction used by the user contains very little communication management. Instead we have annotated utterances as Action-Directive or Report-Task (fail) when they are task-related. Allen and Core (1997) also annotated Communication management, but since we are also interested in utterances related to miscommunication (Green et al., 2006) we have annotated repairs using the categories Repair-Action and Request-repair. Utterances that are aimed at self-repair or manage the speakers contributions have been annotated as Own Communication Management (Allwood et al., 1991).

Perception, Attention and Contact: We have classified contributions related to the management of attention and willingness to interact with the categories Request-and Provide-Attention, and Request- and Provide Contact. Management of contact and attention can be performed using different modalities (Allwood et al., 1991). This draws on findings by (Allwood et al., 1991) and extends the schema adopted by (Gill et al., 2000) who annotated the body move category Attempt-Contact.

The category Signal non-perception (SNP) is similar to the DAMSL category Signal non-understanding (SNU), but fo-

	2.56	2.57	2.58	2.59	3.00
user		robot		this is a book	
com-act		req-att		ref, assert	
robot					
com-act					
u-gest			point		
com-act			ref		
u-gaze		robot	object	robot	



Figure 5: Different corpus data visualized as a score annotation similar to what it may look like in the Anvil tool developed by Kipp (2004). Here we have simplified the image to make it appear better in print.

cuses on the (reported) perceptual status of the participant. A typical example found in the corpus is "I cannot see you" uttered by the robot whenever it lost track of the user.

We have annotated sequences when the user is paying close attention by looking at the robot with the category Monitor. By paying attention to the robot, the user displays a basic positive level of willingness to interact.

Reference: We have annotated events of reference using the category Reference, knowing that there probably is a need to refine this further, for instance, (Gill et al., 2000) uses the more restricted type Demonstrative reference (Dem-Ref). But as this only is used for non-verbal referencing we have chosen the less specific category Reference which we aim to analyze in-depth to arrive at more precise scheme at a later stage.

Emotional display: There are very few occurrences of emotional displays in the corpus. We have annotated obvious examples of emotional display when we have deemed them as being relevant for the communication, e.g., user laughing when the robot speaks something that appears as ill-phrased or out of context. Another category that is related to emotions is Emphasis (Emph), i.e., where a gesture is stressing some aspect of a contribution (e.g., protruding finger during pointing at an object). Furthermore, instances of self-touch, e.g., touching the face or lips have been observed and annotated because it may signal the emotional state of the user or be seen as a sign of invasion of personal space (Sommer, 1969).

3.2. Supportive annotations

The *supportive* or *context-providing* annotations form a heterogeneous set of resources that can be used for different purposes during analysis. For instance, our interest related to the spatiality dimension of embodiment make data on positioning and spatial distance important to analyze users movement patterns. Another interest lies in the relation between dialogue acts and physical acts (Traum, 2000) and how we may use them to analyze the possible goals that the user and robot can possess respectively. For this we need to have a scene overview, and be able to determine the intentional attention of users by analyzing their gaze patterns. We have also annotated the general task that is going on at a specific time to be used as general background information

and organization of tasks at a higher level.

Gaze: We have also annotated the general direction of the user's gaze in terms of domain related concepts, i.e., the robot itself, object ("tv", "telephone") and locations ("corner"). We have also noted that the user looks around in the room when looking for something or while thinking. In the gaze-track this has simply been annotated as "room".

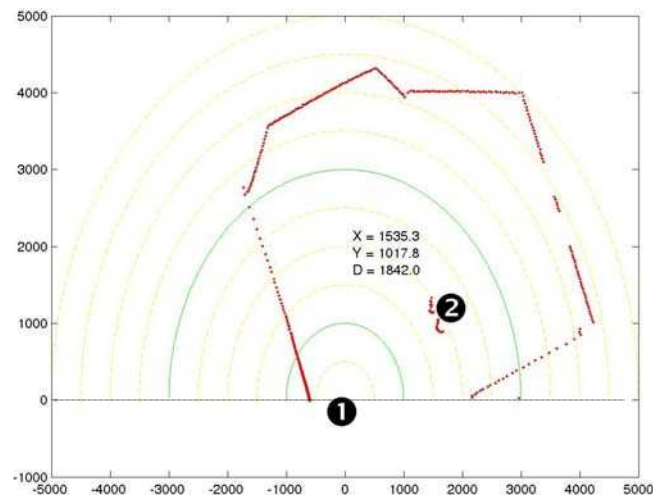


Figure 6: A visualization of laser data using an algorithm for tracking the user (Topp and Christensen, 2005), showing the robot center (1) and the tracked user (2). The walls sensed by the robot are displayed as a (red) dotted line. What appears as two holes in the wall on the right hand side of the image is the shadow of the user's legs (2).

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).

Scenario and users		CORE ANNOTATIONS			CONTEXT/TIME ALIGNED DATA		
		Speech	Comm. acts	Gesture	Task	Posture & Positioning	
Single room	22 users interacting ~15 minutes, task constraints given by system	×	×	×	Objects and locations	Hall distances and F-formations, data from laser range finder	
Multiple room	5 users interacting ~15 minutes, accepting strategy used by the system	×	×	×	Locations stored in system logs		
		MEDIA			BACKGROUND DATA		
		Video	Audio	Webcam	On-board cam	Task-descriptions	Questionnaires/Post session interview
Single room		MiniDV (25 fps)	Stereo onboard (16 kHz), MiniDV (48 kHz)	One web cam in each corner (1 fps)		Time-aligned task annotations	As data file or text document
Multiple room		Handheld MiniDV (25 fps)	MiniDV (48 kHz)		MiniDV (25 fps)		Interview/Videotaped

Table 1: Corpus annotations

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

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. In figure 6 a tracking algorithm has been applied to the data showing the legs of the user as two half circles close to the point indexed (2) in the image.

Task and scene overview: We have annotated tasks on a high level, e.g., as categories related to the general services provided by the system: FOLLOW, SHOW, FIND, GREET, etc. The aim is to provide background information and to visualize organization related to the users' way of solving the task. Another means of providing a general sense of what is going on in the data are images from four network web-cams, that are time aligned to the video. The web-cams were placed in each corner of the single room scenario. It is thus possible to get several perspectives of the scene, and disambiguate the scene linked to the corpus using the timecode. In the multiple room scenario this coverage was not possible to achieve, since it would require a huge amount of cameras. Instead a handheld video-camera provided another perspective on the interaction.

Text descriptions and questionnaire data: During the analysis of spatiality we also wrote down observations on events in the session. These text descriptions have been

time aligned so they can be used as links to specific points of interest in the data. Answers to questionnaires administered to users concerning their attitudes towards the system are also available as a data file.

3.3. 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.

4. 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.

5. References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. webpage. <http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/>.

- Jens Allwood, Joakim Nivre, and Elisabet Ahlsén. 1991. On the semantics and pragmatics of linguistic feedback. Technical Report 64, Gothenburg Papers on Theoretical Linguistics.
- G. Bugmann, S. Lauria, T. Kyriacou, E. Klein, J. Bos, and K. Coventry. 2001. Using Verbal Instruction for Route Learning. In *Proceedings of 3rd British Conference on Autonomous Mobile Robots and Autonomous Systems: Towards Intelligent Mobile Robots (TIMR'2001)*, Manchester, April.
- Guido Bugmann, Ewan Klein, Stanislaw Lauria, and T. Kyriacou. 2004. Corpus-based robotics: A route instruction example. In *Proceedings of IAS-8*, pages 96–103, Amsterdam, NL, March 10–13.
- Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.
- Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies - why and how. *Knowledge-Based Systems*, 6(4):258–256.
- Satinder P. Gill, Masahito Kawamori, Yasuhiro Katagiri, and Atsushi Shimojima. 2000. Role of body moves in dialogue. *International Journal of Language and Communication*, RASK, 12, April.
- Anders Green, Helge Hüttenrauch, and Kerstin Severinson Eklundh. 2004. 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*, pages 575–580, 20–22 Sept.
- Anders Green, Britta Wrede, Kerstin Severinson Eklundh, and Shuyin Li. 2006. Integrating Miscommunication Analysis in the Natural Language Interface Design for a Service Robot. submitted to IROS2006.
- Edward T. Hall. 1966. *The Hidden Dimension: Man's Use of Space in Public and Private*. The Bodley Head Ltd, London, UK.
- Helge Hüttenrauch, Kerstin Severinson Eklundh, Anders Green, and Elin Anna Topp. 2006. Investigating Spatial Relationships in Human-Robot interaction. Submitted to IROS2006.
- Adam Kendon. 1990. *Conducting interaction - Patterns of behavior in focused encounters. Studies in interactional sociolinguistics*. Press syndicate of the University of Cambridge, Cambridge, NY, USA.
- Michael Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Dissertation.com, Boca Raton, Florida.
- M. W. Knudsen, Laila Dykjær, and Niels Ole Bernsen. 2001. 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*, pages 135–146, Aalborg, Denmark, 2 September.
- Y. Koide, T. Kanda, Y. Sumi, K. Kogure, and H. Ishiguro. 2004. An Approach to Integrating an Interactive Guide Robot with Ubiquitous Sensors. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004)*, volume 3, pages 2500–2505, 28 Sept/ 2 Oct.
- Jan Frederik Maas and Britta Wrede. 2006. BITT: A Corpus for Topic Tracking Evaluation on Multimodal Human-Robot-Interaction. In *Proceedings of the Fifth international conference on Language Resources and Evaluation LREC2006*.
- David Mulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an Intelligent Agent through Wizard of Oz. In *INTERCHI'93*, pages 277 – 282. ACM, April.
- Florian Schiel, Silke Steininger, and Ulrich Türk. 2002. The SmartKom Multimodal Corpus at BAS. In *Proceedings of Second International Conference on Language Resources and Evaluation LREC2000*, pages 200–206.
- Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *IUI'04: Proceedings of the 9th international conference on Intelligent User Interfaces*, pages 78–84, New York, NY, USA. ACM Press.
- Robert Sommer. 1969. *Personal Space*. Prentice-Hall.
- Elin A. Topp and Henrik I. Christensen. 2005. Tracking for Following and Passing Persons. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005)*, Edmonton, Alberta, August.
- Elin Anna Topp, Helge Hüttenrauch, Henrik Christensen, and Kerstin Severinson Eklundh. 2006. Acquiring a shared environment representation. In *In Proceedings of HRI2006 1st annual conference on Human-Robot Interaction*, Salt Lake City, UT, USA, March 2–3. ACM.
- David R. Traum. 2000. 20 Questions for Dialogue Act Taxonomies. *Journal of Semantics*, 17(1):7–30.
- Luis Villaseñor, Antonio Mass, and Luis Pineda. 2000. 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, May.
- Joerg C. Wolf and Guido Bugmann. 2005. Multimodal Corpus Collection for the Design of User-Programmable Robots. In *TAROS 2005 Towards Autonomous Robotic Systems Incorporating the Autumn Biro-Net Symposium*, 12th–14th September.

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

Sonja Hüwel, Britta Wrede

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

Abstract

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

1 Introduction

Over the past years interest in mobile robot applications has increased. One aim is to allow for intuitive interaction with a personal robot which is based on the idea that people want to communicate in a natural way (Breazeal et al., 2004)(Dautenhahn, 2004). Although often people use speech as the main modality, they tend to revert to additional modalities such as gestures and mimics in face-to-face situations. Also, they refer to objects

in the physical environment. Furthermore, speech, gestures and information of the environment are used in combination in instructions for the robot. When participants perceive a shared environment and act in it we call this communication “situated” (Milde et al., 1997). In addition to these features that are characteristic for situated communication, situated dialog systems have to deal with several problems caused by spontaneous speech phenomena like ellipses, indirect speech acts or incomplete sentences. Large pauses or breaks occur inside an utterance and people tend to correct themselves. Utterances often do not follow a standard grammar as written text.

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

2 Related Work

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

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

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

There is only little research in semantic speech analysis of spontaneous speech. A widely used approach of interpreting sentences is the idea of case grammar (Bruce, 1975). Each verb has a set of named slots, that can be filled by other slots, typically nouns. Syntactic case information of words inside a sentence marks the semantic roles and thus, the corresponding slots can be filled. Another approach of processing spontaneous speech by using semantic information for the Air Travel Information Service (ATIS) task is implemented in the Phoenix system (Ward, 1994). Slots in frames represent the basic semantic entities known to the system. A parser using semantic grammars maps input onto these frame representations. The idea of our approach is similar to that of the Phoenix system, in that we also use semantic entities for extracting information. Much effort has been made in the field of parsing strategies combined with semantic information. These systems

support preferably task oriented dialog systems, e.g., the ATIS task as in (Popescu et al., 2004) and (Milward, 2000), or virtual world scenarios (Gorniak and Roy, 2005), which do not have to deal with uncertain visual input. The aim of the FrameNet project (Fillmore and Baker, 2001) is to create a lexicon resource for English, where every entry receives a semantic frame description.

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

3 Situated Dialog Corpus

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

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



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

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

4 The Robot Assistant BIRON

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

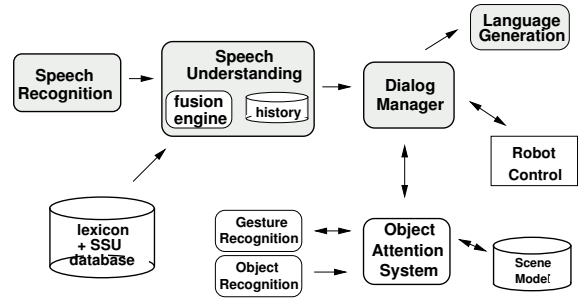


Figure 1: Overview of the BIRON dialog system architecture

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

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

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

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

5 Situated Concept Representations

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

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

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

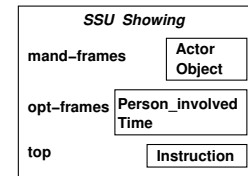


Figure 2: Schematic SSU “Showing” for utterances like “I show you my poster tomorrow”.

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

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

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

matically.

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

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

6 Semantic Processing

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

The fusion mechanism computes in its best case in linear time and in worst case in square time. A scoring function underlies the mechanism: the more words can be combined, the better is the rat-

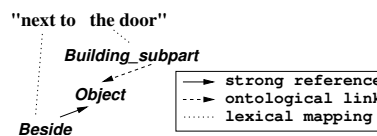


Figure 3: Simplified parse tree example .

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

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

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

```

<metaInfo>
<time>1125573609635</time>
<status>full</status>
</metaInfo>
<semanticInfo>
<u>what can you do</u>
<category>query</category>
<content>
  <unit = Question_action>
    <name>what</name>
    <unit = Action>
      <name>do</name>
      <unit = Ability>
        <name>can</name>
        <unit = Proxy>
          <name>you</name>
        ...
    </unit>
  </unit>
</content>
<u>this is a green cup</u>
<category>description</category>
<content>
  <unit = Existence>
    <name>is</name>
    <unit = Object_kitchen>
      <name>cup</name>
      <unit = Potential_gesture>
        <name>this</name>
      </unit>
    </unit>
    <unit = Color>
      <name>green</name>
    </unit>
  </unit>
  ...
</content>

```

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

6.1 Situated Speech Processing

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

For future work it is planned to integrate additional information sources, e.g., inquiries of the dialog manager to the user. The module will also

```

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

```

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

store these information in the *history* which will be used for anaphora resolution and can also be used to verify the output of the speech recognition.

7 Evaluation

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

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

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

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

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

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

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

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

lead to a failure of the parsing mechanism. 104 utterances have been classified as syntactically not-correct.

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

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

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

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

8 Conclusion and Outlook

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

This mechanism combined with the use of our

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

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

References

- C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda. 2004. Humanoid robots as cooperative partners for people. *Int. Journal of Humanoid Robots*.
- B. Bruce. 1975. Case systems for natural language. *Artificial Intelligence*, 6:327–360.
- K. Dautenhahn. 2004. Robots we like to live with?! - a developmental perspective on a personalized, life-long robot companion. In *Proc. Int. Workshop on Robot and Human Interactive Communication (RO-MAN)*.
- A. Haasch et. al. 2004. BIRON – The Bielefeld Robot Companion. In E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, editors, *Proc. Int. Workshop on Advances in Service Robotics*, pages 27–32. Fraunhofer IRB Verlag.
- C. J. Fillmore and C. F. Baker. 2001. Frame semantics for text understanding. In *Proc. of WordNet and Other Lexical Resources Workshop*. NACCL.
- C. J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conf. on the Origin and Development of Language and Speech*, volume 280, pages 20–32.
- J. Fritsch, M. Kleinhagenbrock, A. Haasch, S. Wrede, and G. Sagerer. 2005. A flexible infrastructure for the development of a robot companion with extensible HRI-capabilities. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 3419–3425.
- P. Gorniak and D. Roy. 2005. Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. In *ICMI*. ACM Press.
- B. J. Grosz and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- H. Hüttenrauch, A. Green, K. Severinson-Eklundh, L. Oestreicher, and M. Norman. 2003. Involving users in the design of a mobile office robot. *IEEE Transactions on Systems, Man and Cybernetics, Part C*.
- S. Hüwel and F. Kummert. 2004. Interpretation of situated human-robot dialogues. In *Proc. of the 7th Annual CLUK*, pages 120–125.
- J. Juster and D. Roy. 2004. Elvis: Situated Speech and Gesture Understanding for a Robotic Chandelier. In *Proc. Int. Conf. Multimodal Interfaces*.
- S. Lauriar, G. Bugmann, T. Kyriacou, J. Bos, and E. Klein. 2001. Personal robot training via natural language instructions. *IEEE Intelligent Systems*, 16:3, pages 38–45.
- S. Li, A. Haasch, B. Wrede, J. Fritsch, and G. Sagerer. 2005. Human-style interaction with a robot for cooperative learning of scene objects. In *Proc. Int. Conf. on Multimodal Interfaces*.
- L. Seabra Lopes, A. Teixeira, M. Quindere, and M. Rodrigues. 2005. From robust spoken language understanding to knowledge acquisition and management. In *EUROSPEECH 2005*.
- J. T. Milde, K. Peters, and S. Strippgen. 1997. Situated communication with robots. In *First Int. Workshop on Human-Computer-Conversation*.
- D. Milward. 2000. Distributing representation for robust interpretation of dialogue utterances. In *ACL*.
- A.-M. Popescu, A. Armanasu, O. Etzioni, D. Ko, and A. Yates. 2004. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In *Proc. of COLING*.
- O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, and R. Dillmann. 2002. Using gesture and speech control for commanding a robot assistant. In *Proc. of the 11th IEEE Int. Workshop on Robot and Human interactive Communication*, pages 454–459. RO-MAN.
- S. Wachsmuth, G. A. Fink, and G. Sagerer. 1998. Integration of parsing and incremental speech recognition. In *EUSIPCO*, volume 1, pages 371–375.
- W. Ward. 1994. Extracting Information From Spontaneous Speech. In *ICRA*, pages 83–86. IEEE Press.
- V. Zue, S. Seneff, J. Glass, J. Polifronti, C. Pao, T. J. Hazen, and L. Hetherington. 2000. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, pages 100–112, January.

Robust Speech Understanding for Multi-Modal Human-Robot Communication

Sonja Hüwel, Britta Wrede, and Gerhard Sagerer
Faculty of Technology, Applied Computer Science
Bielefeld University, 33594 Bielefeld, Germany

shuewel,bwrede,sagerer@techfak.uni-bielefeld.de

Abstract—In order to model complex human robot interaction researchers not only have to consider different tasks but also to handle the complex interplay of different modules of one single robot system. In our context we constructed a robot assistant integrated in a home or office environment. We allow for a fairly natural communication style, which means that the users communicate using speech but are also allowed to use gestures and moreover to use contextual scene knowledge. Against this background, this paper presents a robust speech understanding component for situated human-robot communication. It serves as interface between speech recognition and dialog management. To increase robustness of speech processing it rates the speech recognition output by means of semantic coherence. Even if the recognized word-stream is not grammatically correct the speech understanding component provides semantic interpretations in context of multi-modal input for dialog management. For the understanding process, we designed special semantic concepts grounded to the domain of situated communication. They also provide additional information about the dialog act. A processing mechanism uses these concept units to generate the most likely semantic interpretation of the utterances.

I. INTRODUCTION

Interest in mobile robot applications has increased over the past years. Demands on user-interfaces for such applications are quite different from classical computers, as the use of keyboard or mouse is not appropriate for this type of applications. New technologies oriented on human communication are needed to provide a better way to interact with robots in home or office environments [1], [2]. Naive users can easily interact with these systems without learning a specific programming language. By natural interaction we mean, that the user relies on speech as the main modality, like in the communication between humans. However, natural communication also makes use of other modalities such as gestures [3] and additional cues about the scene or discourse. These additional cues provide important information for the disambiguation and the completion of information from the speech channel. Such a kind of communication is called situated: the participants perceive a shared environment and act in it [4]. Additionally, situated dialog systems have to deal with several problems caused by spontaneous speech phenomena like ellipses. Also, people rarely utter complete sentences as in written text.

Developing a mobile robot system is a complex task. Different aspects as embodiment, personality, and dialog



Fig. 1. Scene from plant-watering task.

have to be respected. Much work has been spent on the design of complex embodied robot systems [5], [6], however, speech understanding for human robot interaction has not yet been extensively explored. Service robots have not only to be able to cope with this special kind of communication but they also have to cope with noise that is produced by their own actuators or the environment. Speech recognition in mobile-robot scenarios is a complex and difficult task, leading to severe degradations of the recognition performance [7]. Robust interpretation of utterances has to consider the specific conditions in HRI. For systems capable of interacting by use of speech and visual input the complexity of speech analysis is mostly reduced to speech recognition [8].

To model complex HRI, we need to consider both: spontaneous speech as well as situatedness. The system ought to allow for a broad variety of utterances which may also be ungrammatical and support multi-modality. Consequentially, the system needs to allow for extralinguistic information as gestures and objects in the environment, use a large lexicon and not restrict the interpretation to a grammar. A widely used approach of interpreting sentences is the idea of case grammar [9]. Each verb has a set of named slots, that can be filled by other slots, typically nouns. Syntactic case information of words inside a sentence marks the semantic roles and thus, the corresponding slots can be filled. However, this approach requires syntactically correct and complete utterances and is difficult to apply to spontaneous speech. An approach of processing spontaneous speech by using semantic information is implemented in the Phoenix

system [10] for the Air Travel Information Service (ATIS) task. Slots in frames represent the basic semantic entities known to the system. A parser using semantic grammars maps input onto these frame representations. However, the task of the system is uni-modal information query, where integration of multiple modalities is not foreseen.

Recent projects with related scope with respect to multi-modality and spontaneous speech are for example the robots CARL [11] and ALBERT [12]. However, as speech recognition in mobile-robot scenarios is a complex and difficult task, in these systems the speech understanding analysis is constrained to a small set of commands and not oriented towards spontaneous speech. However, deep speech understanding is necessary for more complex HRI. Our approach, therefore, is to use semantic entities for extracting information under the specific conditions of HRI.

The paper is organized as follows. First, we will give an overview of the kind of utterances between a human and a robot assistant we have to deal with. In section III, we will briefly describe our robot system BIRON, its overall system architecture and the modules the speech understanding interact with. Then, we will explain the semantic concepts and the mechanism we use to generate appropriate interpretations for HRI. Finally, we will present the evaluation of the entire speech understanding system and end with a conclusion and an outlook for future work.

II. SITUATED DIALOG CORPUS

There is only little research on multi-modal HRI with speech-based robots. A study how users interact with mobile office robots is reported in [13]. However, in this evaluation, the integration of different modalities was not analyzed explicitly. But even though the subjects were not allowed to use speech and gestures in combination, the results support that people tended to communicate in a multi-modal way, nevertheless. To receive more detailed information about the instructions that users are likely to give to an assistant in home or office we simulated this scenario and recorded 14 dialogs from German native speakers. Their task was to instruct the robot to water plants. Since our focus in this stage of the development of our system lies on the situatedness of the conversation, the robot was simply replaced by a human pretending to be a robot. The subjects were asked to act as if it would be a robot. As proposed in [14], a preliminary user study is necessary to reduce the number of repair dialogs between user and system, such as queries. The corpus provides data necessary for the design of the dialog components for multi-modal interaction. We also determined the lexicon and obtained the semantic concepts that describe the scene and tasks for the robot. The recorded dialogs feature the specific nature of dialog situations in multi-modal communication situations. The analysis of the corpus is presented more detailed in [15]. It confirms that spontaneously spoken utterances seldom respect the standard grammar and structure of written sentences. People tend to use short phrases or single words. Large pauses often occur during an utterance

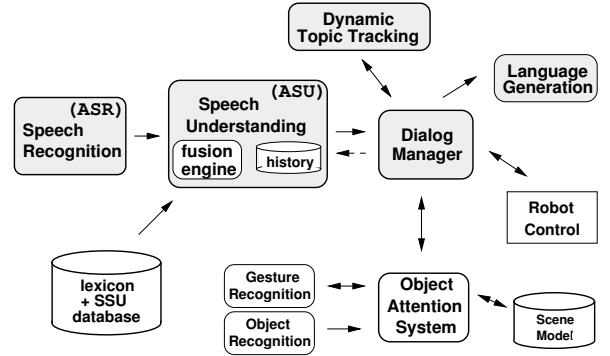


Fig. 2. Overview of the BIRON dialog system architecture

or the utterance is incomplete. More interestingly, the multi-modal data shows that almost every person used pointing gestures in the dialogs to refer to objects. Such utterances cannot be interpreted without additional information of the scene. For example, an utterance such as “this one” is used with a pointing gesture to an object in the environment. We realize, of course, that for more realistic behavior towards a robot a real experiment has to be performed. However this time- and resource-efficient procedure allowed us to build a system capable of facilitating situated communication with a robot. The implemented system has been evaluated with a real robot (see section V).

III. THE ROBOT ASSISTANT BIRON

The aim of our project is to improve social and functional behavior by enabling the robot to carry out a more sophisticated dialog for handling instructions. The basis for this project is the robot system BIRON (Bielefeld Robot Companion) [16]. The robot is supposed to assist the user in home or office environment. First, the user shows the robot its surrounding, including different rooms and important objects, so that it gets to know its own environment. After introducing the robot it can support the user by different tasks, e.g., to water plants, as courier, or to show guests around the place.

The robot BIRON is able to visually track persons and to detect and localize sound sources. The robot expresses its focus of attention by turning the camera into the direction of the person currently speaking. From the orientation of the person’s head it is deduced whether the speaker addresses the robot or not. The main modality of the robot system is speech but the system can also detect gestures and objects. Figure 2 gives an overview of the architecture of BIRON’s multi-modal interaction system. The speech processing parts are grayed. For the communication between these modules we use an XML based communication framework [17]. In the following, we will briefly outline the interacting modules of the entire dialog system with the speech understanding component (see Figure 2).

Speech recognition Initiated by the attention system of the robot the speech recognition system starts to analyze the speech data. For that, we use an incremental speaker-

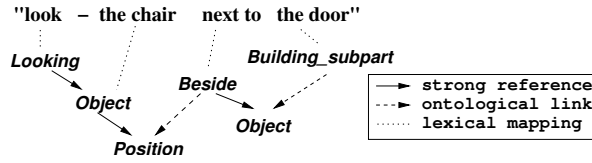


Fig. 3. Simplified parse tree example .

independent speech recognition system [18], based on Hidden Markov Models. It is trained on American English. The system combines statistical and declarative language models to compute the most likely word chain.

Dialog manager The dialog management serves as the interface between speech processing and the robot control system. It also generates answers for the user. Thus, the speech analysis system transforms utterances with respect to gestural and scene information, such as pointing gestures or objects in the environment, into instructions for the robot. The dialog manager in our application is agent-based and enables a multi-modal, mixed initiative interaction style [19]. It is based on semantic entities which reflect the information the user uttered as well as discourse information based on speech-acts. For that, we use discourse segments proposed by Grosz and Sidner [20] to describe the kind of utterances during the interaction. It also needs verbal hints of scene information such as pointing gestures or object descriptions to gather information of the gesture detection and object attention system.

IV. SITUATED CONCEPTS AND SEMANTIC PARSING

Based on the situated conversational data, we designed “situated semantic units” (SSUs) which are suitable for fast and automatic speech understanding. These SSUs basically establish a network of strong and weak relations of semantic concepts which represent world and discourse knowledge. They also provide ontological information and additional structures for the integration of other modalities.

A. Semantic Parsing Algorithm

The database currently contains 1290 lexical entries and 143 SSUs. Each lexicon entry is mapped to an SSU, which may contain strong or weak unresolved references to other SSUs. The semantic processor tries to resolve these open references with SSUs corresponding to other words found in an utterance. Figure 3 shows how the semantic processor handles the utterance “look - the chair next to the door”. In this example “look” is mapped to the SSU *Looking*, representing the root of the semantic structure. “Looking” provides an open reference to the SSU *Object*. The semantic processor scans the word chain for a word with the SSU *Object* or any subtype of *Object* and joins the particle “the chair” to *Object* and thus, to the SSU *Looking*. As objects in the real world are always placed in a certain position, the SSU *Object* has an open reference to *Position*. The lexicon entry “next to” is mapped to the sub-category of *Position* *Beside*, which provides an open reference to the SSU *Object*. The lexicon entry “the door” is mapped to the

SSU *Building_subpart* which is an ontological subtype of *Object*. Thus, “the door” will be linked to the SSU *Beside* corresponding to “next to” which results in a semantic parse tree (Fig. 3).

This example is a simplified illustration of the semantic parsing process. e.g., the words “to” and “the” have separate lexical entries and are processed accordingly. Also the semantic processor computes all possible parse trees for a word chain and rates each one with a scoring value. The scoring function incorporates the amount of combined words, prefers words temporally close-by and distinguishes between strong and weak relations. The parse tree with the highest score is considered to be the most likely one. With this approach variations of the utterance “look at the chair next to the door” can be interpreted. “Look - next to the door - the chair” results in the same semantic structure. The SSU *Looking* provides apart from the *Object* relation a *Position* relation. By this, utterances as “look over there - the plant at the window” can be reduced to a single semantic structure. Further, it is possible to handle homonyms without additional effort, for example in the home-tour scenario the word “can” might be a box or a the ability to do something. This can simply be achieved by two lexicon entries of “can” linked to two different SSUs.

Moreover, the rating increases robustness, it enables to decide whether the speech recognition result is reliable or not. In the case of low score, the dialog manager can ask the user for clarification. The fusion mechanism computes in its best case in linear time and in worst case in square time. The mechanism also uses a *history*, where former parts of utterances are stored and which are also integrated in the fusing mechanism.

B. Dealing with Situated Dialogs

Since we focus on spontaneous speech, we cannot rely on the grammar, and therefore the semantic units serve as the connections between the words in an utterance. If there are open connections interpretable as missing information, it can be inferred what is missing and be integrated by the contextual knowledge. Utterances in this context are normally less complex, still our approach allows a broad variety of utterances. The lexicon and the concept database are inspired by FrameNet [21]. They are based on the experimental data of situated communication (see section II) and also on data of a home-tour scenario with a real robot.

By assigning ontological information, the relations between SSUs can be described as generally as possible, without losing information important for the robot. Disjunction of different subtypes is important for proper robot-actions. The SSU *Building_subpart* for example, a sub-category of *Object*, indicates that this object has a fixed position, which is important for robot navigation-support. This information is stored in the SSU entry *top-cat*. As the dialog manager needs knowledge about discourse segments, the SSUs also reflect these by use of ontological information. For example the SSU *Showing* belongs to the discourse type *Instruction*. Other

```

<metaInfo>
<time>1125573609635</time>
<status>full</status>
</metaInfo>
<semanticInfo>
<u>look here</u>
<category>instruction</category>
<content>
  <unit = "Looking">
    <name>look</name>
    <unit = "Potential_gesture">
      <name>here</name>
    </unit>
  </unit>
  ...
</semanticInfo>
<u>what color is your pencil</u>
<category>query</category>
<content>
  <unit = "Question_attribute">
    <name>what</name>
    <unit = "Attribute">
      <name>color</name>
    </unit>
  <unit = "Existence">
    <name>is</name>
    <unit = "Object_office">
      <name>pencil</name>
      <unit = "Owner">
        <name>your</name>
      </unit>
    </unit>
  </unit>
  ...

```

Fig. 4. Two segments of speech understanding results

types are *Socialization*, *Description* and so on. Using this information enables the dialog manager to react adequately.

In our domain, missing information in an utterance can often be acquired from the scene. For example the utterance “look at this” and a pointing gesture to a table will be merged to the meaning “look at the table”. To resolve this meaning, we also use hints of co-verbal gestures in the utterance. The SSU *Potential_gesture* indicates a relation between speech and gesture. Words as “this one” or “here” indicate the potential existence of a gesture. The timestamp of the utterance enables temporal alignment of the gesture. Since gesture recognition is expensive in computing time and often not well-defined such linguistic hints can reduce these costs dramatically. The utterance “that” can also represent an anaphora, and is analyzed in both ways, as anaphora and as gesture hint. Only if there is no gesture, the dialog manager will decide that the word probably was used in an anaphoric manner.

A converter forwards the result of the mechanism as an XML-structure to the dialog manager. A segment of the result for the dialog manager is presented in Figure 4. With the category-descriptions the dialog-module can react fast on the user’s utterance without any further calculation. It uses them to create inquiries to the user or to send a command to the robot control system, such as “look for a gesture”, “look for a blue object”, or “follow person”. Figure 2 illustrates the entire architecture of the speech understanding system. If the interpreted utterance does not fit to any category it gets the value *fragment*. These utterances are currently interpreted in the same way as partial understandings and the dialog manager asks the user to provide more meaningful information.

C. Examples from real data

In the following we will briefly outline the semantic analysis of real data. The utterance examples in figure 6 are excerpts from the evaluation setting between a human and BIRON (see section V). Pauses are marked with an ‘-’. The utterance of *User1* could be fully analyzed. First, BIRON gets an instruction to look at an object: cup. It also gets information about a potential gesture. Scene and speech input now are joined and stored in the memory. After focusing at the cup by using gestural information about the position, the robot declares, that it found the cup. The same mechanism proceeds with the second utterance. As the user pauses inside the utterance, the speech understanding system uses the *history* information to merge all parts together. The mechanism also ignores the second article “a”. As we didn’t expected meta-comments in this stage, neither the speech recognition nor the speech understanding system was able to interpret the utterance correctly (missing the entry “try”). After modifying the lexicon and SSU database the understanding mechanism now interprets this utterance both as meta-comment as well as “trying x” (while x can be detected by gestural hint). The utterances of *User2* are interpreted similarly.

The first utterance of *User3* is interpreted as the ability to walk in this room. In this case the system does not distinguish between a question and a statement. In both cases, the robot can verify the utterance with its internal knowledge about its abilities and can answer yes or no. As the user also might have asked “you can walk in this room?”, where no syntactic difference to a statement can be found, this approach seems to be suitable. Adding prosodic information might be more helpful for dissolving and could be considered by the scoring function.

In the example of *User4* BIRON answers with “I can move to another location, I can follow you, and if you show me something I can remember it”. Then it starts to dissolve the gestural hint. After *User5* instructed to stop BIRON stops moving (the word “fine” is simply ignored as it is unknown). Then BIRON answers that its name is BIRON. The third utterance could not be recognized properly. The semantic analysis would interpret the utterance only partially, as “got” and “hobbies” are not in the lexicon. It would interpret the utterance as *Possess:Owner:you*.

D. Dynamic Topic Tracking

To improve our system with respect to situation awareness, our robot-system also includes dynamic topic tracking (DTT). Therefore, the speech understanding system allows in combination with the dialog system explicitly uttered topic changes and references (e.g., “let us talk about the kitchen”). The DTT also uses implicit information about topic changes, e.g., utterances as “follow me” or information about the user’s reorientation by looking around, often indicate a topic change. Discourse information and other multi-modal information is also used to topically segment the user utterances. For explicit topic tracking the system provides the SSUs *Topic*, *Theme*, *Change_topic*, *Talk_topic*, and *Question_topic*.



Fig. 5. User interacting with BIRON.

Using these SSUs the system is able to interpret utterances as “what are we talking about” or “let us change the topic” and therefore support situation awareness.

E. Conclusion and Outlook

Our approach has various advantages when dealing with spontaneous speech. The system can process incomplete or grammatically incorrect utterances and furthermore, rate the semantic coherence via the scoring function. Repeated words as in the utterance “look - look here” do not interfere the mechanism. Corrections inside an utterance as “the left - em right cube” are analyzed correctly (preferring the rightmost) and the system also deals with pauses inside utterances by integrating former parts of utterances stored in the *history*. In our system we also use scene information to evaluate the entire correctness so that we do not only have to rely on the speech input. In case of doubt the dialog-manager requests to the user.

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

V. EVALUATION

For the evaluation of the entire robot system BIRON we recruited 14 naive user between 12 and 37 years with the goal to test the intuitiveness and the robustness of all system modules as well as its performance. Therefore, in the first of two runs the users were asked to familiarize themselves with the robot without any further information of the system. In the second run the users were given more information about BIRON’s functions.

Having a closer look at the dataset gathered by the experiment, we observed similar results as described in section II. Users mostly utter short phrases. They tend to pause in the middle of an utterances, especially, when a gesture is involved, as in “look - this - is a cube”. They also often uttered so called meta-comments as “oh - actually”, “that’s fine” or “let’s try that one”. As the subjects were non-native English speaker, they also commentated in their native language German, which caused severe speech recognition errors. Many users also changed their behavior, depending on how much BIRON could understand. First, they talked to it in a normal way. Repeating an utterances the first time,

```
User1: Look here robot - a cup.
      Now look here a - a keyboard.
      Let's try that one. ...
User2: Robot look - do you see?
      This - is a cow. Funny.
      Do you like it? ...
User3: Can you walk in this room?
      How fast can you move? ...
User4: What can you do?
      Oh - so what is that? ....
User5: Please stop here - stop moving - fine.
      What is your name?
      Have you got any hobbies? ...
```

Fig. 6. Excerpts from user utterances during the experiment.

they hyper-articulate words and also speak louder. Some of them switched to baby-talk style, e.g., “BIRON look - do you see - this - is a cow. Fine. Do you like it?”. In figure 6 some excerpts of the dialogs during the experiment settings are presented.

Thus, not surprisingly the speech recognition error rate in the first run was 60% which decreased in the second run to 42%, with an average of 52%. This high error rate was also a result of our experiment setting as we wanted to test the intuitiveness and robustness of the entire System. As we also wanted to evaluate the intuitiveness of the entire system, subjects were not informed about the vocabulary and the kind of utterances the system can understand. To avoid this high error rate in future we will improve BIRON’s personal character to make more visible the internal states and of the robot and its abilities. But in the case of intuitive and complex interaction the high error rate also proves the need of robust speech understanding.

As to the speech understanding part, we wanted to get information about its robustness especially in the way it interprets the speech recognition output: the correctly recognized utterances and also the misrecognized utterances. The semantic speech processing received 1642 utterances from the speech recognition system. In total the interaction time was 151 minutes with an average of 5 minutes 14 seconds per user. The speech understanding system processed all utterances and performed on a standard PC with an average processing time of 20ms, which fully fulfills the requirements of real-time applications. As speech recognition not always guarantees correct results especially with spontaneous speech, the understanding mechanism is underlying a rating function. 22% of the results were rated as complete misunderstandings, 17% as partial understandings or could not be assigned to a discourse pattern (mostly single words) and therefore not be processed by the dialog manager, and 61% as correct utterances with full semantic meaning. Thus, the speech understanding component was able to recover substantially from the speech recognition errors and also to improve robustness. Comparing the error rate of the speech recognizer with the rating output of the speech understanding system (see table I) we can detect that the system was able to reduce the high error rate and moreover to decide between senseful and senseless speech recognition output. Only 4% of the utterances which were correctly recognized

	ASR		ASU
false	52%	not interpretable (non + partial)	39%
correct	48%	interpreted (full) by dialog manager	61%

TABLE I
SPEECH RECOGNITION ERROR RATE COMPARED TO SCORING OF
SEMANTIC INTERPRETATIONS.

were misinterpreted or refused by the speech understanding system. Most errors occurred due to missing words in the lexicon. We will also use these results to verify and adjust the generated structures. and thereby improve the SSUs.

On the other hand, many utterances which were only recognized partially correct still could be interpreted semantically and this also holds true for ungrammatical utterances. For example users uttered syntactically incorrect utterances as “look here BIRON look left” (interpreted as “look: (potential_gesture (person: BIRON) (orientation: left))”), “a book this is”, or “look cow”. These utterances still could be transformed into a unique semantic meaning. There are many utterances which were slightly misrecognized as one word is misunderstood or noise of the environment is interpreted as extra speech input. This kind of speech recognition output could also be transformed into uniform semantic structures. Thus, the speech understanding component was able to recover substantially from the speech recognition errors and also to improve robustness (see table I).

VI. CONCLUSION AND OUTLOOK

In this paper we have presented a new approach of robust speech understanding for mobile robot assistants. This mechanism combined with the use of SSUs has several benefits. First, speech is interpreted even if speech recognition does not always guarantee correct results and speech input is not always grammatically correct. Secondly, the speech understanding component incorporates information about gestures and references to the environment. Furthermore, the mechanism itself is domain-independent: both, concepts and lexicon can be exchanged. This semantic analysis already produces elaborated interpretations of utterances in a fast way and furthermore, helps to improve robustness of the entire speech processing system.

We are now planning new experiments to evaluate the understanding mechanism without limitation by the speech recognition system. Our goal is to evaluate the entire variety of utterances unexperienced users may express to BIRON. Additionally, the dynamic topic tracking will be tested within the setting. The user may utter explicit topic changes, i.e., “let us talk about food”. We will also use the evaluation results to improve the SSUs to get better results for the semantic interpretation and to prevent false connections.

ACKNOWLEDGEMENTS

This work has been supported by the European Union within the “Cognitive Robot Companion” (COGNIRON)

project (FP6-IST-002020) and by the German Research Foundation (DFG) within the Graduate Program ‘Task Oriented Communication’.

REFERENCES

- [1] 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.
- [2] K. Dautenhahn, “Robots we like to live with?! -a developmental perspective on a personalized, life-long robot companion,” in *Proc. Int. Workshop on Robot and Human Interactive Communication (RO-MAN)*, 2004.
- [3] D. McNeill, *Hand and Mind: What Gesture Reveal about Thought*. University of Chicago Press, 1992.
- [4] J. T. Milde, K. Peters, and S. Strippgen, “Situating communication with robots,” in *First Int. Workshop on Human-Computer-Conversation*, 1997.
- [5] B. J. A. Kröse, J. M. Porta, A. J. N. van Breemen, K. Crucq, M. Nuttin, and E. Demeester, “Lino, the User-Interface Robot,” in *European Symposium on Ambient Intelligence (EUSAI)*, 2003, pp. 264–274.
- [6] N. Tomatis, R. Philippsen, B. Jensen, K. O. Arras, G. Terrien, R. Piguet, and R. Siegwart, “Building a Fully Autonomous Tour Guide Robot: Where Academic Research Meets Industry,” in *Proc. Int. Symp. on Robotics*, 2002.
- [7] G. A. Fink and S. Hohenner, “Experiments in distant talking speech recognition using a standard database,” in *Proc. 31. DAGA '05*, 2005.
- [8] H. G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano, “Human-robot interaction through real-time auditory and visual multiple-talker tracking,” in *Proc. of the 2001 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2001.
- [9] B. Bruce, “Case systems for natural language,” *Artificial Intelligence*, vol. 6, pp. 327–360, 1975.
- [10] W. Ward, “Extracting Information From Spontaneous Speech,” in *Proc. IEEE Int. Conf. on Robotics and Automation*, 1994, pp. 83–86.
- [11] L. S. Lopes, A. Teixeira, M. Quindere, and M. Rodrigues, “From robust spoken language understanding to knowledge acquisition and management,” in *EUROSPEECH 2005*, 2005.
- [12] O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, and R. Dillmann, “Using gesture and speech control for commanding a robot assistant,” in *ROMAN*, 2002, pp. 454–459.
- [13] H. Hüttenrauch, A. Green, K. Severinson-Eklundh, L. Oestreicher, and M. Norman, “Involving users in the design of a mobile office robot,” *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 2003.
- [14] S. Lauriar, G. Bugmann, T. Kyriacou, J. Bos, and E. Klein, “Personal robot training via natural language instructions,” *IEEE Intelligent Systems*, 16:3, pp. 38–45, 2001.
- [15] S. Hüwel and F. Kummert, “Interpretation of situated human-robot dialogues,” in *Proc. of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, 2004, pp. 120–125.
- [16] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinhagenbrock, S. Lang, I. Topsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, “BIRON – The Bielefeld Robot Companion,” in *Proc. Int. Workshop on Advances in Service Robotics*, E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, Eds., 2004, pp. 27–32.
- [17] J. Fritsch, M. Kleinhagenbrock, A. Haasch, S. Wrede, and G. Sagerer, “A flexible infrastructure for the development of a robot companion with extensible HRI-capabilities,” in *Proc. IEEE Int. Conf. on Robotics and Automation*, Barcelona, Spain, April 2005, pp. 3419–3425.
- [18] S. Wachsmuth, G. A. Fink, and G. Sagerer, “Integration of parsing and incremental speech recognition,” in *Proc. of the European Signal Processing Conference*, vol. 1, 1998, pp. 371–375.
- [19] 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*, 2005.
- [20] B. J. Grosz and C. L. Sidner, “Attention, intention, and the structure of discourse,” *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.
- [21] C. J. Fillmore and C. F. Baker, “Frame semantics for text understanding,” in *Proc. of WordNet and Other Lexical Resources Workshop*, 2001.

1 Research Interests

My research interest lies in **multi-modal dialog management systems** for **human-robot interaction**. More specifically, I am interested in multi-modal dialog management systems for **mobile, personal service robots with learning capabilities** that work at home of humans.

1.1 Past, Current and Future Work

I am working for the EU-funded project COGNIRON (Cognitive Robot Companion). The goal of this project is to develop autonomous, personal robots that can increase their knowledge about the *real-world* by interacting with humans. For example, in one of our key-experiments, the *home-tour* the user shows a newly purchased robot her home to prepare it for future tasks. When the user says “This is my computer” and points to it, the robot should be able to link the verbally specified name of the object with “real-world features” of the object. In particular, it should (1) understand user’s intention with the symbolic name “computer”, (2) detects the object computer in the current scene according to her gesture, (3) add this object into its spatial representation of the current room and (4) store all the multi-modal information about the computer in its memory. The focus of my work is to develop a dialog management system for such interactions.

I proposed a *computational model of multi-modal grounding* (Li, 06) that is able to flexibly handle human-robot dialog and integrate multi-modal information systematically into the dialog process. For the basic grounding model I combined the advantages of the previous work (Traum, 94), (Cahn and Brennan, 99) and introduced the notion of *grounding relation* which enables me to model the grounding process using an augmented push-down automaton. I represent interlocutors’ contributions as exchanges, i.e., pairs of contributions. They achieve the state “grounded” only if the acceptance of the presentation is available which depends on the communication success and the robot task execution status. These exchanges are organized in a stack and the grounding status of the whole stack is dependent on the status of the individual exchanges and the grounding relations between them. I introduced 4 types of such relations (*default*, *support*, *correct* and *delete*) and they can also have

local effects on their previous exchanges. Each contribution of both the user and the robot is categorized in terms of its roles, 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”. To systematically integrate multi-modality processing into the dialog I proposed to represent the agents’ dialog contribution as a two-layered structure: the *interaction unit*. On the domain layer a communicational intention is conceived and on the conversation layer language in verbal and non-verbal form is generated according to the communicational intention. The domain layer is the interface of agents to their back-end applications (or to her cognitive capacity in case of human agent). During the dialog, the robot agent studies the verbal and non-verbal information contained in the conversational layer of the user’s interaction unit and then transfers this information to the robot control system by creating an interaction unit for its own. After receiving the response of the robot control system the robot agent conceives a communication intention on its domain layer and generates verbal and non-verbal output. Any interpretation and execution failure in this process results in the creation of new exchanges with appropriate grounding relations to the current exchange.

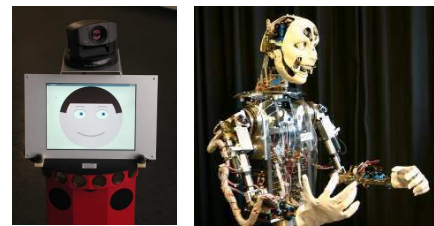


Figure 1: Robots BIRON and BARTHOC

This model was implemented on the mobile robot BIRON and ported to the humanoid robot BARTHOC (Fig. 1). The details of the realization of multi-modal handling is described in (Li, 05). My next step is to realize automatic user adaptation for the dialog management system to increase the usability of our robots.

2 Applications and Challenges in Spoken Dialog Systems Research

The next killer application is mobile, personal robots with learning abilities. Such robots differ from other applications in several ways: Firstly, the issue of knowledge. In most dialog applications the system possesses more information than the user (e.g., train schedule applications) or has the same mental concept as the user (e.g., cooperative planning systems). In the home-tour scenario, however, the roles of the user and the system are reversed: the robot has to build mental concepts only according to what the user says or does. Secondly, The issue of multi-modality. The multi-modality of human-robot interaction is different from that in applications like multi-modal PDAs in that the robot has to observe human non-verbal behavior *unobtrusively* and infer their meanings in the current conversational context. Thirdly, the issue of situatedness. A robot is situated in the real world and has to be able to handle real objects and qualities. Concepts like form, color, weight, distance, etc. can not be only realized on the symbolic level, but need to be grounded like in humans. This point differentiates robots from most virtual agents. All these three issues effects the dialog modeling strategies.

Impact of this application: A mobile personal robot with learning abilities can be viewed as the most intelligent computer applications because it could adapt itself both to the environment and to its user. It could learn where is the living room, learn to cook spaghetti and remember the football match schedule. And the best thing is, if the user moved into a new house or fell in love with Chinese food she did not need to re-program the robot, she just needed to say “Look, BIRON, this is my new living room, and please serve dumpling for lunch.”

Questions that need to be answered: Firstly, the separation of dialog planning from task planning. It is essential to enable the dialog system for a learning robot to handle dialog without much domain knowledge. It should be possible to at least externally configure the interface to the domain task execution and the response generator so that it can be easily updated or replaced. Secondly, the question of usability. A robot is a highly complex system with a large number of modules working cooperatively. Technical necessities can sometimes result in serious usability problems. For example, it is necessary for BIRON to shut down its speech recognizer if it can not “see” the user’s face because it should only process speech if there is a dialog partner around. This means if the face detector fails BIRON will not “listen to” the user at all. For the user, BIRON just stops to react for unknown reasons. Such “chain reactions” are not rare and can become serious usability problems. Therefore, a compromise has to be found between technics and usability.

ity. Thirdly, the question of evaluation. Many modules of a robot system run complex algorithms, claim plenty of computation time and are subject to environmental conditions, e.g., lighting conditions, noise, size of doors, obstacles etc. For the dialog system, this means that the correct interpretation and transaction of user utterances is by no means a guaranty for a prompt response or successful task execution. Thus, the performance of the dialog system can not be directly measured with the performance of the overall system. New evaluation methodologies have to be defined for human-robot dialog systems.

3 Suggestions for Discussion

- Advantages and limitations of Wizard of Oz studies.
- Advantages and challenges of multi-modal dialog systems that incorporate vision-based technologies
- How to bridge the gap between dialog research in different application fields

References

- S. Li, B. Wrede, and G. Sagerer 2006. *A computational model of multi-modal grounding for human-robot interaction*, in Proc. 7th. SIGdial Workshop on Discourse and Dialogue.
- S. Li, A. Haasch, B. Wrede, J. Fritsch, and G. Sagerer 2005. *Human-style interaction with a robot for cooperative learning of scene objects*, Proc. Int. Conf. on Multimodal Interfaces.
- D. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D Thesis, University of Rochester.
- J. E. Cahn and S. E. Brennan. 1999. *A Psychological Model of Grounding and Repair in Dialog*, Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems.

Biographical Sketch



Shuyin Li is a Ph.D student working in the Group of Applied Computer Science at the Bielefeld University in Germany. She received her Master Degree in Information Processing at the University of Cologne, Germany (2003) and her Bachelor Degree in German Language and Culture at the University Qingdao, China (1998). From 1997 to 2004 she received three awards from the Bavarian State Ministry of Science, Research and the Arts (Germany), Siemens Corporation and German Academic Exchange Service, respectively.

Why and how to model multi-modal interaction for a mobile robot companion

Shuyin Li and Britta Wrede

Applied Computer Science, Faculty of Technology
Bielefeld University,
33594 Bielefeld, Germany
Email: shuyinli, bwrede@techfak.uni-bielefeld.de

Abstract

Verbal and non-verbal interaction capabilities for robots are often studied isolated from each other in current research trend because they largely contribute to different aspects of interaction. For a robot companion that needs to be both useful and social, however, these capabilities have to be considered in a unified, complex interaction context. In this paper we present two case studies in such a context that clearly reveal the strengths and limitations of these modalities and advocate their complementary benefits for human-robot interaction. Motivated by this evidence we propose a powerful interaction framework which accounts for flexible multi-modality management in complex interactions.

Introduction

Humanoid robots are viewed as unique platforms for developing human-like capabilities for machines. Based on their appearance and their hardware equipment these robots are likely to exhibit believable, human-like behaviors which facilitate a smooth and intuitive interaction with human users. At present, two main research strands can be observed in the field of interaction study. The first strand emphasizes the emotional and social expressivity of *non-verbal behaviors* and targets at the modeling of human-like behaviors for robots whose sole purpose is to engage human users in social interaction (Breazeal 2002) (Kanda *et al.* 2002). Their evaluation focus is often how enjoyable or human-like the interaction is for the user. The second strand concerns the development of *conversational* systems for robots (Matsui *et al.* 1999), (Lemon *et al.* 2001), that provide services for human users and communicate with them mainly using speech. In their evaluations mainly the efficiency and success ratio of the (mostly) task-related interaction are tested. Although the issues of these two strands contribute both to the human-robot interaction (HRI) research they are largely studied in an isolated manner.

The focus of our work is to develop an interaction framework for our robot companion BIRON (The Biele-

feld Robot Companion). A robot companion is a robot that should be able to both accomplish useful tasks and behave socially (Dautenhahn *et al.* 2005). It is thus essential for us to take into account both task-related and social factors of interaction which requires a combination of the two research strands. To achieve this goal we first have to clarify the advantages and disadvantages of verbal and non-verbal modalities in a complex interaction context so that we can use the appropriate modality for different purposes. Researchers in the first strand tend to argue for the general importance of non-verbal behaviors. For example in case of facial expressions, it is argued in (Breazeal 2002) that human-like expressions are easy to understand for human users. The author of (Billard 2005) holds the view that a face can give user clues as to the ability of the robot. For example, a two-year-old face implies less cognitive capability than a adult face. In contrast, researchers in the second strand emphasize more the importance of speech which is, almost indisputable, the main channel to communicate task-related information between the user and the robot. In our view, the issue of strengths and limitations of modalities should not be discussed in an isolated context. Instead, we should draw conclusions from real, complex human-robot interaction where both task-related and social aspects are manifested. More specifically, the robot can be configured to communicate these aspects *only* in either verbal or non-verbal modality because this may result in an suboptimal interaction which reveals limitations of the involved modality and also possible complementary advantages of other modalities. In this paper we report our observations made during a user study with BIRON which, at the time of the study, could only generate verbal feedback. From these observations we draw conclusions clearly in favor of developing additional non-verbal feedback capabilities for BIRON because of benefits that can not be derived only by speech.

Having realized the importance of both verbal and non-verbal robot behaviors in HRI we need an interaction framework that can handle these behaviors in a flexible way. Researchers in the field of virtual agents have been developing such systems since decades and have proposed many sophisticated ap-

proaches which, similar as in HRI, can be classified in two categories. The first category focuses on the optimal animation of synchronized, multi-modal behaviors for virtual characters in *individual utterances* and advocate an relatively static architecture-based approach. Based on (Thórisson 1997) and her own earlier work, Cassell (Cassell 2000) proposed a generic architecture for virtual agents and implemented it for REA (real estate agent) (Cassell *et al.* 2000). In this architecture, the Input Manager collects input from all modalities. Data that requires instant reaction is sent via Hardwired Reaction to the Action Scheduler that generates multi-modal output in a synchronized manner. Input data that needs deliberate discourse processing is forwarded to the deliberative module for interpretation and response generation. The second category concerns approaches that focus on the *whole process* of communication with multi-modality as different forms of contribution. Traum (Traum & Rickel 2002) proposed a multi-modal conversational model based on the *information state theory* (Larsson & Traum 2000) which is a powerful mechanism to model complex conversation. His model consists of 11 layers and sublayers representing different aspects of conversation and the state of these layers can be changed by dialog acts in verbal or non-verbal form. This model, thus, accounts for multi-modality in a very complex dialog context. In this paper we present a multi-modal interaction framework which attempts to avoid the problems and combine the advantages of the both approaches above.

The rest of the paper is organized as follows: section reports the observations we made during a user study with BIRON and section presents our multi-modal interaction framework. Finally, section summarizes the necessity of incorporating verbal and non-verbal interaction capabilities into HRI and the strength of our approach.

Observations

Our robot BIRON is a personal robot with social learning abilities. It can detect and follow persons, focus on objects (according to detected deictic gestures of the human user), and store collected multi-modal information into a memory. Our implementation scenario is the so-called *home tour*: a user bought a new robot from a shop and shows it her home to prepare it for future tasks. The robot should be able to learn and remember features of objects that the user mentions and it “sees”, e.g., name, color, images etc. For example, if the robot learns how the favourite cup of the user looks like and where it is, it can fetch a cup of tea with it for the user later. This scenario is also a learning scenario for the user because she interacts with a complex robot system for the first time. Lacking knowledge about its capabilities the user has to totally rely on the robot’s ability to explain things itself.

The user study we conducted was originally intended to test various features of two new modules that we integrated into the system. Although the interaction

system with the capability to handle input and output of different modalities was already integrated, the robot face, which was intended to be the main non-verbal feedback media, was not yet completed so that it was excluded from the study. We recruited 14 subjects aged from 20 to 37 from the Bielefeld University. None of the subjects had interacted with BIRON before. They were just told to try out BIRON’s functionalities without any knowledge about BIRON’s technical limitations. With this setting we intended to simulate the situation of the “innocent user” in the home tour. In average, each subject interacted with BIRON more than 7 minutes. After the interaction the subjects were asked to fill out a questionnaire. The quantitative results of this study are discussed in (Li, Wrede, & Sagerer 2006b). For the purpose of this paper, we only focus on the observations that we made as a third person and report two cases that illustrate the limitation of the speech modality that can be compensated by non-verbal modalities.

Case one: meta-commentators

The interaction system of BIRON is able to assess the performance of the robot in the runtime and take initiative (approximately every four user-robot-exchanges) to generate output to praise or comfort the user according to the assessment results. This feature was designed to show users some level of social awareness of BIRON which should motivate them and also make the interaction more enjoyable.

During the experiment we observed differences in interaction styles among the subjects which result in their different reactions to this feature of the system. Our scenario of home tour has an explorative nature for subjects because they do not have much pre-knowledge about the robot’s capabilities, including its language capabilities. Therefore, most subjects started the interaction by saying things to see the reaction of BIRON which often caused out-of-vocabulary errors in speech recognition at the beginning of the interaction. The interaction system of BIRON then asked the user to rephrase or try again and also, from time to time, comfort her with utterances like “I know it is sometimes difficult with me, but don’t feel discouraged!”. This behavior turned out to be an effective means to reduce frustration level for the majority of subjects (as revealed by the questionnaires (Li, Wrede, & Sagerer 2006b)). However, as we observed, it is also this behavior that caused more frustration in some other users. These were “meta-commentators” who tended to reply to such apologies by saying, e.g., “I’m already discouraged.” or “Apology has no use.” etc. These comments were most times out of vocabulary of the robot which caused even more speech recognition errors and apologies from the robot. The upshot of this spiral of robot speech was, from the perspective of the meta-commentators, it apologized for its performance far too often and the whole interaction entered into a dead-lock around the apologies.

Now we seem to be in a dilemma: the robot’s com-

ments on its own performance are basically a good idea but for some users they cause more trouble than generate benefits. Since we have to account for different users we need to find out what of the robot should be improved to solve this problem. We can identify two features of the speech modality, which we used to convey the social awareness, as possible reasons for the problem: reciprocity and obtrusiveness. Verbal conversation is a social act and is subject to social conventions. If one utters something the conversation partner usually has the obligation to reply (Clark 1992), often using the same modality. Since the robot issued the apologies verbally it is natural for the subjects to also reply verbally. For us, however, it is difficult to predict what should be included into the robot's vocabulary because the content of the reply is highly person-dependent. Additionally, speech is obtrusive in general because the listener is more or less "forced" to listen to it given the conversational convention of turn-taking. In human-machine interaction it is also often the case that the machine is not able to execute other tasks before the current issue is resolved. Thus, the meta-commentators had to repeatedly listen to the apologies and barging in caused even more speech recognition errors. The solution to the problem, therefore, lies in replacing the verbal apologies with non-verbal ones.

The strength of non-verbal feedback is its unobtrusiveness because users can decide themselves whether or not she pays attention to these behaviors. Besides, non-verbal behaviors generally do not impose so strong obligations to reply on the interaction partner as speech so that the potential for unexpected replies is reduced. Additionally, non-verbal feedback can be demonstrated in parallel to speech which enables a fluent task-oriented conversation without losing social capabilities of the robot. In our case what can help is, e.g., a robot face that is able to demonstrate apologizing facial expressions and grows more and more sorrowful while the interaction quality gets worse. At a certain point speech can also be involved to make the concern of the robot more explicit. Thus, the robot is able to regulate its social-aware behaviors according to the severity of the situation.

Case two: quiet speakers

As a *situated* computer device (Brooks 1989) a robot can not take it for granted that the user is permanently present and willing to interact with it. We therefore developed a Person Tracking and Attention system (PTA) for BIRON (Lang *et al.* 2003) that can detect humans and interpret their interaction intention by searching and analyzing the (combination of) perceived human legs, voice location and face direction. The speech recognizer is activated only if the PTA identifies the interaction intention of a human user, i.e., if the status of the the robot system is changed from "Person detected" to "Communication Partner (CP) detected". This policy makes sense because it prevents the robot from erroneously attempting to interact with the user if the user

is talking with another person or it is just the radio on. For the interaction system of BIRON, however, this means, several requirements have to be met before the processing of user speech begins: (1) the person has a reasonable distance to the robot; (2) the person stands straightly; (3) the person is looking at the robot's camera; (4) the person speaks clearly and reasonably loud and (5) the robot correctly receives and analyzes all the incoming signals. In the real system operations these requirements sometimes may not be fulfilled which has serious consequences on the interaction.

Our speech recognizer is trained with male voices and, therefore, does not work particularly well with female voices. One of our female subjects spoke with very low voice in high pitch so that the speech recognizer most of the time interpreted her voice as noise and did not forward it for further processing. Puzzled by no reaction from the robot, the subject looked frequently in the direction of the experimentator and asked why the robot did not react. Knowing that her own voice could also influence the robot's perception of the environment the experimentator tried to use gesture to make clear that she could not intervene. The subject seemed not to be able to interpret the meaning of experimentator's gesture and, therefore, went a few steps towards the experimentator so that she was out of the range where the robot could perceive her as a human. Then the subject came back and tried again, in vain. In this whole process, although the robot went through different internal states (Person detected, CP detected, CP lost, Person lost, Person detected, CP detected), there was no visible reaction from the robot for the subject which was a very frustrating interaction experience for her.

The problem of this case is that we had no means to convey the internal states to the subject before the active interaction starts. With our only available modality speech we only had two possibilities. The first one is to make the robot to report its internal state every time when it is changed. However, this can mislead users to believe that a connection to the robot is already established which is not true because these states only represent the robot's perception and a connection is only established when the robot receives explicit speech from the user. The second possibility can avoid this problem by giving users the impression that the speech output on the state changes is some routine output from the robot: it can repeatedly generate output like "I saw you, I saw you, I saw you...". This is obviously not really an option with the speech modality, but a good solution if the robot had non-verbal feedback capabilities. Non-verbal behaviors can be effectively used to represent static information which is occasionally updated. In our case, these behaviors can help users to understand how the robot perceives a human and, if needed, to adjust her own behaviors, e.g., not walking out of the view of the robot.

Speech can convey complex semantic information and is therefore indispensable for both human users and

robots to communicate in a task-related context. However, as our case studies show this modality reaches its limit when the robot also has to simultaneously demonstrate its internal awareness of human user’s possible affective status and physical presence. Thus, for a robot companion that should be able to carry out useful tasks and behave socially, both verbal and non-verbal modalities have to be included into its interaction framework.

A multi-modal interaction framework

In this section, we propose an interaction framework that is able to handle multi-modal input and output in a flexible way. This framework is based on a computational model of multi-modal grounding (Li, Wrede, & Sagerer 2006a) and is designed for face-to-face human interaction with robot companions. The central idea of our approach is to represent individual interaction contributions as Interaction Unit and organize them based on the principle of grounding. In this section, subsection first discusses problems of conventional approaches, subsection describes our framework of multi-modal interaction and subsection summarizes the benefits of this framework.

The background

Interaction systems operate with so-called *contributions*, i.e., the input from the user and output to the user which both can change the state of the system and probably trigger further actions. Most systems represent these contributions directly with user/robot utterances or their non-verbal behaviors. The assumption of this view on contribution is that each input or output in whatever modality conveys a meaning that should be expressed. This assumption has two problems. Firstly, it is difficult to explain the role of simultaneous verbal and non-verbal contributions because the relationship between the meaning they convey is variable. For example, Scherer (Scherer & Wallbott 1979) stated that non-verbal behavior can substitute, amplify, contradict and modify the meaning of the verbal message. Iverson et al. (Iverson *et al.* 1999) studied human gestures and identified three types of *informational* relationship between speech and gesture: reinforcement (gesture reinforces the message conveyed in speech, e.g., emphatic gesture), disambiguation (gesture serves as the precise referent of the speech, e.g., deictic gesture accompanying the utterance “this cup”), and adding-information (e.g., saying “The ball is so big.” and shaping the size with hands). Taking each input or output as the contribution can hardly account for these various relations between the verbal and non-verbal messages. Secondly, treating verbal and non-verbal input and output as equal contributions means that they have to share the same semantics. However, some modalities may primarily convey semantic information, e.g., speech, while some other modalities more affective information (e.g., facial expression). They can hardly share the same semantics and it is even more difficult when they happen simultaneously. Most current studies bypass these

problems by adding modules that accomplish the additional analysis which result in, in our view, a heterogeneous structure of the framework.

The interaction framework

Rather than directly taking verbal and non-verbal behaviors as individual contributions we go one level deeper and propose *interaction intention* as the basic contribution. Note, we adopt a relatively broad definition of intention here: *motivation of behaviors that can evoke interaction behaviors in other persons*. According to this definition also emotions that are manifested in some way can be viewed as an interaction intention if other persons react to it. For example, if Tom looks sad and Jane asks “Are you OK?”, then Tom’s sadness is also an interaction intention. Of course, Tom may not intend to communicate his sadness originally, but the act of interaction between Tom and Jane is already established. Thus in this case, we also view Tom’s sadness as interaction intention. As to the relations between intention and the manifested interaction behavior we draw upon findings of cognitive science on the process of multi-modal language production. Levelt (Levelt 1989) identified three phases of language production: *conceptualization*, *formulation*, and *articulation*. The production of an utterance starts from the conception of an *intention* and the semantic organization in the conceptualization phase, and then 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 is to say that the intention is the driving force of interaction. Evidence for the assumption that the verbal and non-verbal conversational behaviors are two surface forms of the underlying intention is delivered by McNeill (McNeill 1992), who stated that the production of speech and conversational gesture is governed by one single “idea unit”. These findings motivate us to set up a two-layered structure to represent contributions: Interaction Unit (IU).

Each IU consists of one *Intention Layer* and one *Conversation Layer*. On the Intention Layer (ILayer), an intention as defined above is conceived. For a robot the intentions come from the analysis of the previous discourse or from the other modules of the robot system. On the Conversation Layer (CLayer), this intention is formulated and articulated¹. Two generators (verbal and non-verbal generator) on this layer are responsible for generating spoken language and non-verbal behavior according to the intention conceived, respectively. These two layers represent the intention conception and the multi-modal language generation process, respectively. The structure of the IU is illustrated in Fig. 1.

According to this model, interaction partners participate in an interaction by generating such IUs and reply

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

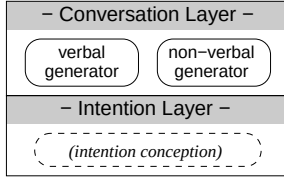


Figure 1: The structure of an Interaction Unit (IU)

to IUs of the partner. Since one either initiates an account or replies to the account of the other during an interaction their intentions can be either *self-motivated* or *other-motivated*. Both the user and the robot can generate self-motivated IUs when she/it feels it needs to be. For a robot, self-motivated intentions can be triggered by other modules of the robot system, e.g., perception modules that have just detected a human presence. In this case, an IU can be created with its ILayer importing the message from the robot module and exporting an intention. This intention is transferred to the CLayer which then formulates a verbal message with the verbal generator and/or constructs a non-verbal expression with the non-verbal generator, e.g., saying “Hello, human” and smiling. Note, the verbal and non-verbal generator do not always need to be simultaneously instantiated. It depends on the initiator of the IU to choose the appropriate modalities to express the intention. Other-motivated intentions can be triggered by the needs of the on-going interaction, e.g., the need to answer a question, or be triggered by the robot’s execution results of the task specified previously by the user. When a robot replies to a human user’s IU, as illustrated by the UML activity diagram in Fig. 2, it first has to understand the human user’s intention by studying the CLayer of her IU. Here it only studies the instantiated generator, i.e., if user speech is perceived which means the verbal generator is instantiated, then this generator has to be studied; if both generators are instantiated, both have to be studied. The goal of this study is to figure out the underlying intention of the user. If the user intention can not be recognized the robot should clarify this. This clarification intention is then used by the robot to build an IU that, e.g., verbally utters “pardon?” using its verbal generator and/or constructing a questioning face with its non-verbal generator. If the intention can be recognized the robot can go on to transfer the user intention to the responsible module of the robot system and wait for its response. Once the response is available, the interaction system can proceed to create its own IU with the intention that is conceived based on this response.

These IUs have to be put into the joint interaction discourse which records the content of the interaction so far and regulates the interaction with rules. We employ the principles of grounding that were first proposed by Clark (Clark 1992) to organize the discourse. For the reason of space we only briefly summarize its essentials here (for details see (Li, Wrede, & Sagerer 2006a)). The

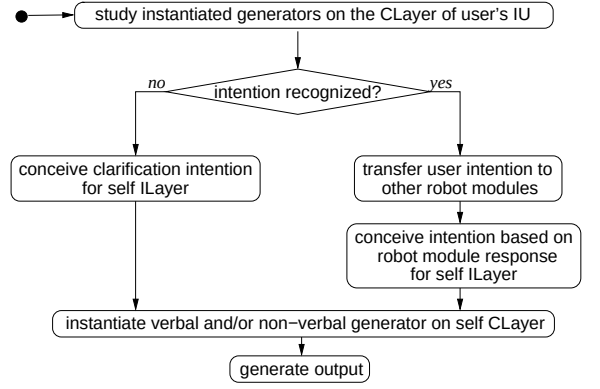


Figure 2: UML diagram for the processing steps of an other-motivated IU

notion of grounding states that during a conversation the agents need to coordinate their mental states based on their mutual understanding about the current intentions, goals and tasks. He termed this process “grounding”. Furthermore, a speaker can only be sure that her account (Presentation) was fully understood if her conversation partner provides some evidence of understanding (Acceptance), i.e., if the “common ground” is available. Only then will the speaker be willing to proceed to another account. Since face-to-face HRI bears the essential features of a conversation (namely reciprocity), we extend this idea of grounding to account for broader interactions. In our system we represent a pair of contributions, i.e., a pair of IUs, as an exchange. These exchanges achieve 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. Acceptance can be implicit or explicit which is a modification of Clark’s *strength of evidence principle* and addresses the phenomena that not all the Presentations have to be explicitly accepted using speech. 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 introduce four types of such relations (*default*, *support*, *correct* and *delete*) and they can also have local effects on their previous exchanges. According to the execution results of the robot system the interaction system creates contributions for the robot. Each contribution of both the user and the robot is categorized in terms of its roles, 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 into a vector which records the complete interaction history. The structure of the interaction discourse is illus-

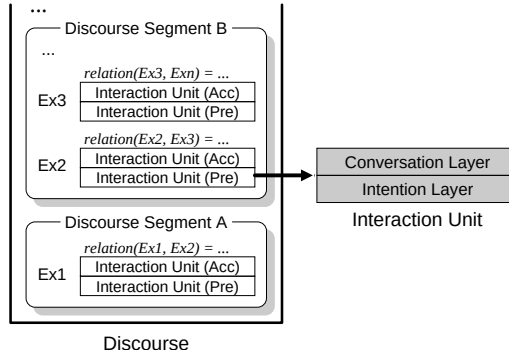


Figure 3: The structure of the interaction discourse (Ex: exchange; Pre: Presentation; Acc: Acceptance)

trated in Fig 3. We thus model the grounding process using an augmented push-down automaton which exhibits local flexibility. The implemented system enables a mixed-initiative interaction style and a well-organized discourse maintaining mechanism. It can also handle complex conversational repair behavior which facilitates a smooth interaction.

This framework was already implemented on our mobile robot BIRON and humanoid robot BARTHOC (Fig. 4) that use this mechanism to interpret user conversational gestures (Li *et al.* 2005). Our user study (Li, Wrede, & Sagerer 2006b) also provides strong evidence for the flexible interaction style enabled by this framework. In the current system, the two interaction cases discussed in section are handled as following: For the case of meta-commentators, the system generates an IU with only the non-verbal generator on the CLayer being instantiated to express its regret for doing a bad job. What the user sees is thus only a sorrowful face from the robot. This IU initiates an Exchange that only requires an implicit Acceptance which means that this Exchange is still grounded even if no Acceptance from the user is perceived. For the case of quiet speakers, the user’s physical presence is viewed as a non-verbal Presentation of the user, for which the system should provide an Acceptance. The system does this by generating an IU that performs a certain facial expression on its CLayer representing its internal state. When the status of the user is changed from “Person” to “Communication Partner”, i.e., when the user proposes a new Presentation, the system generates a new IU as Acceptance, on the CLayer of which a new facial expression is then presented. Thus, our framework covers a broader range of interaction which distinguishes itself from most other systems that can only account for active interaction between the user and the robot.

Potentials of the framework

In general, our framework combines the advantages of the two directions in interaction modeling as discussed in section : the representation of contributions as IUs enables a synchronized generation of verbal and non-

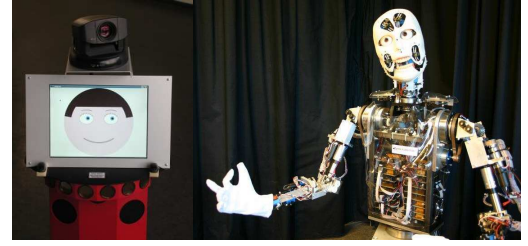


Figure 4: Robots BIRON and BARTHOC

verbal output and organizing IUs based on the grounding principle allows flexible management of the interaction discourse. This framework can also be extended in two ways. Firstly, although in our work so far we focused on the disambiguation function of gesture to speech which requires only a relatively simple modality fusion mechanism, a “Modality Manager” can be added into the CLayer that accomplishes sophisticated modality fusion and selection in the language understanding and generation phase, respectively. It can also be granted access to the interaction discourse to facilitate its analysis. This local extension will not affect the overall interaction management. Secondly, emotion processing can be integrated into this framework in two different ways. One possibility is to view it as one criterion for modality fusion and selection in the Modality Manager which means that the emotion arises basically from the intention that is conceived on the ILayer. The other possibility is to integrate an independent module of Emotional Engine, the state of which also affects the intention conception process in the interaction system. Last but not least, the IUs can also be used to extend interaction systems with a simple discourse management mechanism (e.g., finite state-based) since it only affects the representation of local interaction contributions. Since the majority of the current conversational interaction systems running on robots are finite state-based (Li, Wrede, & Sagerer 2006a) this concept of IUs would enable many other robots to handle multi-modal input and output without changing the underlying interaction management mechanism.

Conclusion

In this paper we discussed the importance of taking into account both verbal and non-verbal behaviors for a robot by drawing upon our observations of two interaction cases in a user study. These observations confirmed the complementary benefits of these behaviors for HRI. We also proposed a powerful interaction framework that combines advantages of different conventional approaches and can flexibly handle complex, multi-modal human-robot interaction. After we complete the construction of facial expressions for our robots we will conduct new user studies to further verify our framework.

Acknowledgment

This work is supported by the European Union within the Cognitive Robot Companion (COGNIRON) project (FP6-002020)

References

- Billard, A. 2005. Challenges in designing the body and the mind of an interactive robot. In *AISB05, Symposium on Robot Companions: Hard Problems and Open Challenges*.
- Breazeal, C. L. 2002. *Designing Sociable Robots*. MIT Press.
- Brooks, A. R. 1989. A robot that walks; emergent behaviors from a carefully evolved network. MIT AI Lab Memo 1091.
- Cassell, J.; Bickmore, T.; Campbell, L.; and Vilhjalmsson, H. 2000. Human conversation as a system framework: Designing embodied conversational agents. In Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E., eds., *Embodied conversational agents*. MIT Press.
- Cassell, J. 2000. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM* 43(4).
- Clark, H. H., ed. 1992. *Arenas of Language Use*. University of Chicago Press.
- Dautenhahn, K.; Woods, S.; Kaouri, C.; Walters, M.; Koay, K. L.; and Werry, I. 2005. What is a robot companion - friend, assistant or butler?
- Iverson, J. M.; Capirci, O.; Longobardi, E.; and Caselli, M. C. 1999. Gesturing in mother-child interactions. *Cognitive Development* 14(1):57–75.
- Kanda, T.; Ishiguro, H.; Ono, T.; Imai, M.; and Nakatsu, R. 2002. Development and evaluation of an interactive humanoid robot robovie. In *Proc. Int. Conference on Robotics and Automation*.
- Lang, S.; Kleinhagenbrock, M.; Hohenner, S.; Fritsch, J.; Fink, G. A.; and Sagerer, G. 2003. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proc. Int. Conf. on Multimodal Interfaces*.
- Larsson, S., and Traum, D. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering* 6(3–4).
- Lemon, O.; Bracy, A.; Gruenstein, A.; and Peters, S. 2001. Information states in a multi-modal dialogue system for human-robot conversation. In *Proceedings Bi-Dialog, 5th Workshop on Formal Semantics and Pragmatics of Dialogue*.
- Levelt, W. 1989. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Li, S.; Haasch, A.; Wrede, B.; Fritsch, J.; and Sagerer, G. 2005. Human-style interaction with a robot for cooperative learning of scene objects. In *Proc. Int. Conf. on Multimodal Interfaces*. ACM Press.
- Li, S.; Wrede, B.; and Sagerer, G. 2006a. A computational model of multi-modal grounding. In *Proc. SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL*. ACL Press.
- Li, S.; Wrede, B.; and Sagerer, G. 2006b. A dialog system for comparative user studies on robot verbal behavior. In *Proc. 15th Int. Symposium on Robot and Human Interactive Communication*. IEEE Press.
- Matsui, T.; Asoh, H.; Fry, J.; Motomura, Y.; Asano, F.; Kurita, T.; Hara, I.; and Otsu, N. 1999. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services.
- McNeill, D. 1992. *Hand and Mind: What Gesture Reveal about Thought*. University of Chicago Press.
- Scherer, K. R., and Wallbott, H. G., eds. 1979. *Die Funktionen des nonverbalen Verhaltens im Gespräch, in Nonverbale Kommunikation: Forschungsberichte zum Interaktionsverhalten*. Beltz Verlag.
- Thórisson, K. R. 1997. Gandalf: An embodied humanoid capable of real-time multimodal dialogue with people. In *Proc. 1st ACM International Conference on Autonomous Agents*.
- Traum, D., and Rickel, J. 2002. Embodied agents for multi-party dialogue in immersive virtual world. In *Proc. 1st Int. Conf on Autonomous Agents and Multi-agent Systems*.

A computational model of multi-modal grounding for human robot interaction

Shuyin Li, Britta Wrede, and Gerhard Sagerer

Applied Computer Science, Faculty of Technology

Bielefeld University, 33594 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, handle multi-modal information involved in the communication and be relatively independent of the domain knowledge. Most dialog systems developed for mobile robots today, however, are often system-oriented and have limited capabilities. We present an agent-based dialog model that are specially designed for human-robot interaction and provide evidence for its efficiency with our implemented system.

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 that 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 affects the dialog modeling. Here we want to first clarify the most essential requirements for dialog management 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 perceived 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.

The second challenge for HRI dialog management is the *embodiment* 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 to both dialog partners, the speaker tends to say “this cup” while pointing to it. The same strategy is considerably ineffective during a phone call. This example shows, an HRI dialog system must account for multi-modal communication.

The third, probably the unique challenge for HRI dialog management is the implication of the learning ability of such a robot. Since a personal service robot is intended to help human in their individual household it is impossible to hard-code all the knowledge it will need into the system, e.g., where the cup is and what should be served for lunch. Thus, it is essential for such a robot to be able to learn new knowledge and tasks. This ability, however, has the implication for the dialog system that it can not rely on comprehensive, hard-coded knowledge to do dialog planning. Instead, it must be designed in a way that it has a loose relationship with the domain knowledge.

Many dialog modeling approaches already exist. McTear (2002) classified them into three main types: *finite state-based*, *frame-based*, and *agent-based*. In the first two approaches the dialog structure is closely coupled with pre-defined task steps and can therefore only handle well-structured tasks for which one-side led dialog styles are sufficient. In the agent-based approach, the com-

munication is viewed as a *collaboration between two intelligent agents*. Different approaches inspired by psychology and linguistics are in use within this category. 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 conversational agent that performs communicative acts. During a conversation, the dialog system selects the communicative goal based on its current belief about the domain and general conversational obligations. Such systems make use of communication and domain model to enable mixed-initiative dialog style and to handle more complex tasks. In the HRI field, due to the complexity of the overall systems, usually the finite-state-based strategy is employed (Matsui et al., 1999; Bischoff and Graefe, 2002; Aoyama and Shimomura, 2005). As to the issue of multi-modality, one strand of the research concerns the fusion and representation of multi-modal information such as (Pfleger et al., 2003) and the other strand focuses on the generalisation of human-like conversational behaviors for virtual agents. In this strand, Cassell (2000) proposes a general architecture for multi-modal conversation and Traum (2002) extends his information-state based dialog model by adding more conversational layers to account for multi-modality.

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 extension of this model to handle multi-modal grounding. In section 3 we outline the capabilities of the implemented system and present some quantitative evaluation results.

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 displaying a 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. In this section, we also put

concrete examples from the robot domain to clarify the relatively abstract model.

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 applied to model dialog. The first one is the recursivity of Acceptance. Clark claimed, since everything said by one 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 may never really end (Traum, 1994). The second critical issue is taking contributions as the most basic *grounding units*. 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 utterance 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 explicitly commenting Tom’s name, yet the three exchanges that Tom initiated were all accepted.

Organization of grounding units: In accordance with 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 individual 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 ex-

change 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 up to the current state are no longer relevant and the agents do not need to ground them any more.

Note, once an exchange is grounded it is *immediately* removed from the stack so that its IPE becomes the IPE of the next exchange. This model is described as an augmented push-down automaton (Fig. 2). It is augmented in so far that transitions can trigger actions and a 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 the 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 (e.g., “defaultP” stands for “Default Presentation”) and Acceptance.

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

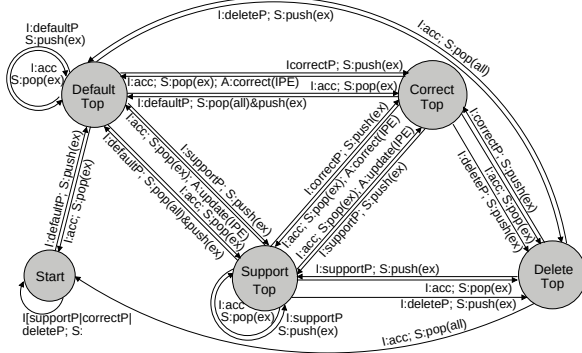


Figure 2: Augmented push-down automaton for grounding (ex: exchange)

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 its validity is deleted. For the reason of space, we only 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 initiates a new exchange, say Ex2, with grounding relation Support to Ex1 and is pushed onto the stack. Then contribution A2 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 grounds 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. For an implemented system, this means that both the user and the system can 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 here¹. 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 language (e.g., gesture) can happen simultaneously. McNeill (1992) stated that gesture and speech arise from the same semantic source, the

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

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 *informational* relationship between speech and gesture:

reinforcement (gesture reinforces the message conveyed in speech, e.g., emphatic gesture), *disambiguation* (gesture serves as the precise referent of the speech, e.g., deictic gesture accompanying the utterance “this cup”), and *adding-information* (e.g., saying “The ball is so big.” and shaping the size with hands). In our work, when processing users’ multi-modal contributions we focus on the disambiguation relation; when creating multi-modal contributions for the robot we are also interested in other informational relations². The structure of an IU is illustrated in Fig. 3.

Operation flow within an interaction unit:

During a conversation an agent either initiates an account or replies 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 transferred 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 robot’s execution results of the tasks specified previously by the user. The operation flow is similar to that of

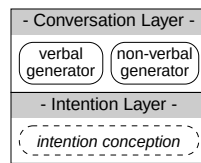


Figure 3: IU

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 dialog. Although it is not the focus of our work, our model can also handle purely non-verbal contributions, since the verbal generator does not always need to be activated if the non-verbal generator already provides enough information about the speaker’s intention. Possible scenarios are: the user looks tired (presentation) and the robot offers “I can do that for you.” (acceptance) or the user says something (presentation) and robot nods (acceptance).

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.,

²This policy has a practical reason: it is much more difficult in computer science to correctly recognize and interpret human motion than to simulate it.

the human and the robot work towards the same goal. With this theory in mind we describe how we process other-motivated contributions below.

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. Since we focus on disambiguation function of non-verbal behavior 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.

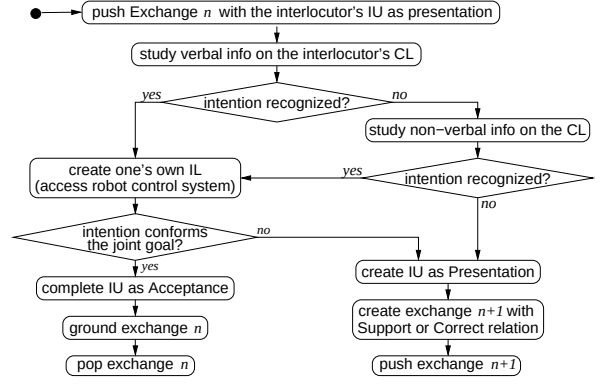


Figure 4: Handling other-motivated contribution (CL: Conversation layer; IL: Intention Layer)

What the dialog system needs to know from the robot control system is what processing results it can produce. The association of these results with robot intentions in terms of whether they start a new account, support or correct one, or delete it, can be configured externally and thus easily updated or replaced. Based on this configuration 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 human deictic gestures) and store collected 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 and remember features of objects that the user mentions and it “sees”, e.g., name, color, images etc. Besides, our system was also successfully ported to a humanoid robot BARTHOC for studies of emotional and social factors of HRI (see. Fig. 5).

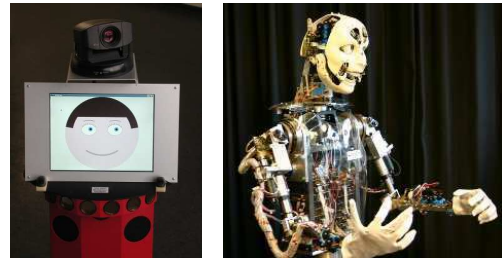


Figure 5: Robots BIRON and BARTHOC

The dialog manager is linked to a speech understanding system which transforms parts of speech

from a speech recognizer into 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, viewed as user's Acceptance of exchanges that the robot initiates. The main part of the dialog system is the Dialog Manager that carries out 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 repair behaviors. With the example dialog in Fig. 6, which is transcribed from a real interaction session between a user and BIRON, we discuss the two most important features of our system.

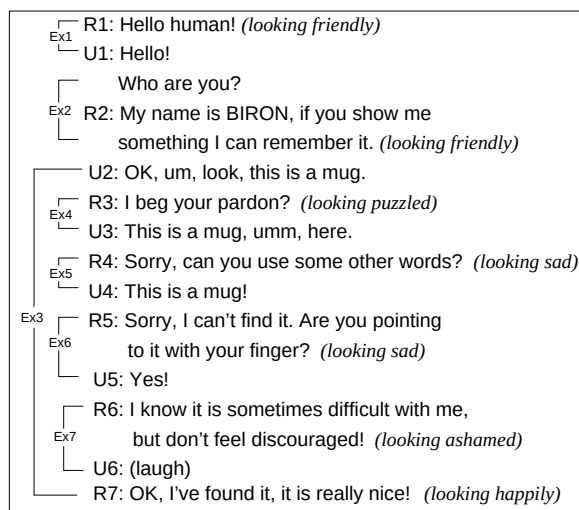


Figure 6: A dialog example with the extrovert BIRON. (U: user, R: robot, Ex: Exchange)

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 it (additionally to its *basic* personality) that takes communication-related initiatives. The basic BIRON behaves in a rather passive way and only says something when addressed by the user. In contrast, the extrovert BIRON greets persons actively (R1 in Table 6) and re-

marks on its own performance (R6). When the robot control system detects a person 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 the robot's 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 z". 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-R7 in Table 6) 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 vision 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 does not oppose (R5-R7). This process and the associated multi-modality fusion and representation are described in (Li et al., 2005) in detail.

The evaluation of dialog systems for human robot interaction is still an open issue. A robot system is usually a complex system including a

large number of modules that claim plenty of processing time or are subject to environmental conditions. For the dialog system, this means that the correct interpretation and transaction of user utterances is by no means a guaranty for a prompt response or successful task execution. Thus, the performance of the dialog system can not be directly measured with the performance of the overall system like most desktop dialog applications. We are still working at evaluation metrics for HRI dialog systems (Green et al., 2006). But the efficiency of our system is already visible in the small effort associated with the porting of this system to another robot platform and in the pilot user study with BIRON. In this study, each of the 14 users interacted with BIRON twice. In the total 28 runs the dialog system generated 903 exchanges for the 813 user utterances. Among these exchanges, 34% initiated clarification questions. This result correlated with the evaluation result of our speech understanding system which fully understood 65% of all the user utterances. 18.6% of the exchanges were Support exchanges created due to execution failure of the robot control system which corresponds to the performance of the robot control system. The average processing time of the dialog system was 11 msec.

4 Conclusion

In this paper we presented an agent-based dialog model for HRI. The implemented system enables multi-modal, mixed-initiative dialog style and is relatively domain independent. The real-time testing of the system proves its efficiency. We will work out detailed evaluation metrics for our system to be able to draw more general conclusion about the strength and weakness of our model.

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*.
- 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*.
- J. Cassell, T. Bickmore, L. Campbell, and H. Vilhjalmsson. 2000. Human conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied conversational agents*. MIT Press.
- H. H. Clark, editor. 1992. *Arenas of Language Use*. University of Chicago Press.
- A. Green, K. Severinson-Eklundh, B. Wrede, and S. Li. 2006. Integrating miscommunication analysis in natural language interface design for a service robot. In *Proc. Int. Conf. on Intelligent Robots and Systems*. submitted.
- J. M. Iverson, O. Capirci, E. Longobardi, and M. C. Caselli. 1999. Gesturing in mother-child interactions. *Cognitive Development*, 14(1):57–75.
- 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*.
- T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu. 1999. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services,. In *Proc. AAAI Nat. Conf. and Innovative Applications of Artificial Intelligence Conf.*
- 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*.
- N. Pflieger, J. Alexandersson, and T. Becker. 2003. A robust and generic discourse model for multimodal dialogue. In *Proc. 3rd Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.
- E. A. Schegloff and H. Sacks. 1973. Opening up closings. *Semiotica*, pages 289–327.
- D. Traum and J. Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual world. In *Proc. 1st Int. Conf on Autonomous Agents and Multi-agent Systems*.
- D. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

A dialog system for comparative user studies on robot verbal behavior

Shuyin Li, Britta Wrede, and Gerhard Sagerer
Applied Computer Science, Faculty of Technology
Bielefeld University, 33594 Bielefeld, Germany
Email: shuyinli, bwrede, sagerer@techfak.uni-bielefeld.de

Abstract—In domestic social robot systems the dialog system is often the main user interface. The verbal behavior of such a robot, therefore, plays crucial role in human-robot interaction. Comparative user studies on various verbal behaviors of a robot can effectively contribute to human-robot interaction research. In this paper we present a dialog system that can be easily configured to demonstrate different verbal, initiative-taking behaviors for a robot and, thus, can be used as a platform for such comparative user studies. The pilot study we conducted does not only provide strong evidence for this suitability, but also reveals benefits of comparative studies on a real robot in general.

I. INTRODUCTION

Robotics researchers are becoming increasingly interested in user-centered design issues of human-robot interaction (HRI). Especially for robots designed to work with non-professional users in home environments many comparative studies have been conducted to find out user perception of robots and user preferences. Goetz et al. [1] investigated the relevance of robot appearance and behavior to user acceptance; Woods et al. [2] studied the relationship between user personality and the perceived robot personality; Walters et al. [3] explored the socially acceptable spatial distance between human users and a robot. In such comparative studies the robots were mostly controlled by a Wizard of Oz (WoZ) remote controller. Other researchers are more interested in real robot testings to find out improvement potentials of the given system. Severinson et al. [4] discussed social aspects of HRI departing from their experience with a fetch-and-carry robot for a motion-impaired user in an office environment. Kanda et al. [5] reported their 2-weeks field study of two interactive humanoid robots with children. Both WoZ and real robot testing provide valuable insight into human behavior in HRI.

However, there is only few studies on effects of robots' verbal behavior, neither in form of WoZ nor real robot testing. Although in many WoZ studies speech is used to demonstrate the behavior to be investigated (e.g., [1] uses different dialog patterns to produce playful and serious robot behaviors that the authors intended to test), crucial verbal behavior issues like how and when the robot should take initiative, what the robot should say to produce certain effects are seldom the focus of the study in HRI. This does not mean, that these issues are not relevant. On the contrary, for many domestic social robots the dialog system is the main interface between the user and the robot. The robot verbal behavior

realized by a dialog system is explicit and clearly observable. Using speech, the intention of the robot can be conveyed in a direct way so that human users can directly evaluate it. Comparative user studies that expose subjects to various verbal behaviors of a robot are, therefore, an effective way to study user perception and preferences in HRI. Based on our own experience with system development and user studies [6] we attribute this “unpopularity” of comparative studies on robot verbal behavior to their methodological challenges:

Firstly, the content of verbal messages is relevant in such a study. The most important task of a dialog system is to communicate information about the robot's internal status to the user. Particularly in situations like robot technical failure the user can only rely on the dialog system to learn what has happened. The content of this information, especially the information about the current system status, is crucial in a user study on robot verbal behavior. However, in a WoZ study it is difficult to simulate all or many of the factors that affect the performance of the robot especially because many of these factors also have cross-correlations, e.g., the performance can be influenced by environmental conditions which in turn can also be affected by the user herself.

Secondly, the way when and how verbal messages are conveyed is also relevant. A dialog system directly communicates with users and they often infer the overall behavior of a robot from the behavior of the dialog system. For example, a robot that frequently takes task-related initiative may be perceived as more intelligent by users. This means, when and how the dialog system “says” something is an important factor in user studies on verbal behavior because they may shape robot's personality or its ability of social awareness. Using real robot systems to study such aspects requires a dialog system that can (1) demonstrate behaviors that reasonably affect user's impression in some way and (2) be able to demonstrate different behaviors to facilitate the comparative nature of such studies. However, most dialog systems implemented for robots today are based on *finite-state based approach*, e.g., [7], [8], [9], [10]. In this approach, the dialog system follows a script of prompts that are pre-defined for the domain tasks and the users need to answer these questions to make the dialog system to reach its goal state. Although it is still possible to use such systems to generate output of different verbal styles, it is difficult to vary “deeper” dialog aspects like initiative distribution that is not confined to individual utterances in certain context but

changes the whole interaction behavior of the robot. Thus, it is difficult to use these systems to study issues like when should the robot say something, which have great influence on the quality of *situated* communication like HRI.

Having looked at the special challenges of comparative studies on effects of robot verbal behavior we are convinced that we should combine the advantages of the two user study approaches for our robot. This means, *to develop a real dialog system with various verbal behaviors*. In user studies with such a system, different verbal behaviors can be directly contrasted to each other that are based on the identical, realistic, technical environment. Additionally to gaining insight related to verbal behavior, new research issues could also arise from this combination of the two user study methods (as discussed at the end of section III).

In this paper, we present such a dialog system for a mobile social robot BIRON (The Bielefeld Robot Companion) that can be easily configured to demonstrate different initiative taking behavior. After describing the dialog management mechanism in section II we forward to section III to discuss the results of a pilot study that provides strong evidence of the suitability of our system for conducting comparative user studies on robot verbal behavior. In this section we will also discuss the benefits we derive from this study. Finally, in section IV, we summarize the strength of our approach and outline potential use of this system.

II. THE DIALOG SYSTEM

In general, a dialog system is responsible for carrying out verbal interaction with users including transferring user commands to the robot control system and reporting task execution results to the user. During the conversation a dialog system should be able to regulate the initiative distribution, handle miscommunication, draw inferences between interlocutors' contributions and organize and maintain the discourse. To enable these abilities we implemented a powerful *agent-based* dialog model for our system. The basic idea of this approach is to view dialog as a collaboration between two equally intelligent agents both of whom take part in the conversation in the same way (see [11] for an overview).

A. The Dialog Management Mechanism

One of the most influential theories on the collaborative nature of dialog is the common ground theory of Clark [12]. In his opinion, during a conversation the interlocutors need to coordinate their mental states based on their mutual understanding about the current intentions, goals, and tasks. He terms this process as "grounding". Furthermore, a speaker can only be sure that her account, the so-called *presentation*, was fully understood if her interlocutor provides some evidence of understanding, the so-called *acceptance*, this is to say, if the "common ground" is available. Only then will the speaker be willing to proceed to another account.

We extended this grounding idea for robot applications: We represent interlocutors' contributions as *Interaction Units* (IU). Each IU has two layers: the intention layer and the conversation layer. On the intention layer a communication

intention is conceived which is usually based on the status messages from the underlying robot system. On the conversation layer a verbal or non-verbal output is generated based on the intention created on the intention layer. During a dialog, the IUs of the interlocutors mostly occur as pairs, e.g., one asks a question and the other answers it. We call these pairs of IUs *exchanges*. In each exchange, one IU plays the role of the presentation and the other the acceptance. Such an exchange is only grounded if the accepting IU is available. Sometimes new exchanges need to be created to ground the preceding exchange, e.g., if one says something that the other does not understand. In this case the listener probably initiates a new exchange to clarify what the speaker said. This means, there can be *grounding relations* between exchanges. We introduced 4 types of such relations (*Default*, *Support*, *Correct*, and *Delete*). These relations can have local effects on their preceding exchanges. For the reason of relevance we only discuss the relation *Default* and *Support* in more detail:

Default: The current Presentation introduces a new account that is independent of the previous exchange in terms of grounding. For example, if an agent says "Hello, my name is Tom, what is your name?" this utterance is segmented into three Presentations which 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. Examples of such exchanges are Ex4, Ex5, and Ex6 in the dialog segment illustrated in Fig. 1. They are all initiated to support the grounding process of Ex3. If such a *Support* exchange is grounded its initiator will try to ground the preceding exchange again with the newly collected information through the supporting exchange. For our robot BIRON, we define, it may create acceptance IUs, i.e., to ground an exchange initiated by the user, only if (1) it understands the user utterance and (2) it successfully executes the task specified by the user. Otherwise BIRON has to initiate a *Support* exchange to let the user rephrase or say something else.

We organize these exchanges into a stack which represents the whole ungrounded discourse. If one interlocutor introduces an account into the discourse, one exchange is pushed onto the stack; if one exchange is grounded, then it is popped from the stack. According to different grounding relations actions can occur when an exchange is pushed or popped. Besides, the discourse is divided into discourse segments according to topics. Figure 2 illustrates the organization of the discourse. In short, we model the grounding process using an augmented push-down automaton which exhibits local flexibility in contrast to conventional grounding approaches ([13], [14]). Our system enables mixed-initiative dialog style and can handle complex conversational repair. (Our system is also able to handle multi-modal IUs, see [15] for more information. A detailed description of this dialog model can be found in [16].)

For the purpose of HRI user studies, the probably most outstanding feature of this system is that it is easy to make

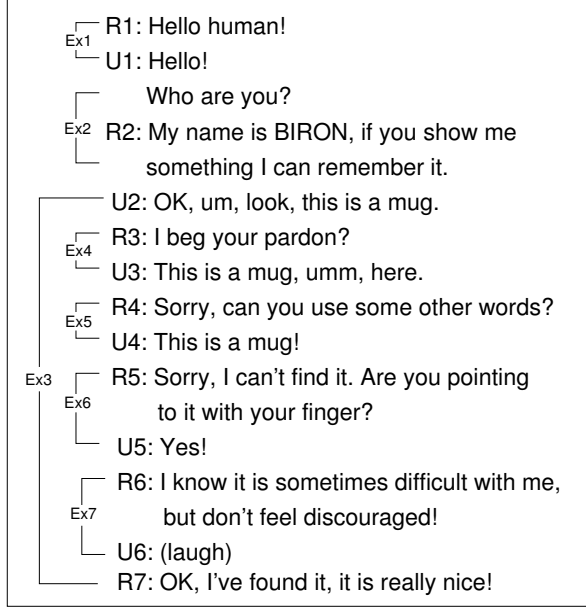


Fig. 1. Taking initiatives (Ex: exchange, U: user, R: robot)

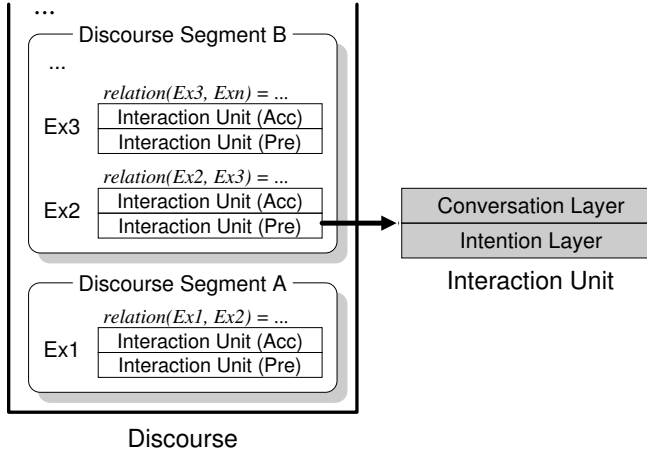


Fig. 2. The structure of the discourse (Ex: exchange; Pre: presentation; Acc: Acceptance)

the system “say something” and to handle user’s response to it from the discourse point of view. In the spirit of the agent-based dialog modeling approach, the domain tasks, in our case the status messages from the underlying robot system, are viewed as motivations for the system to create contributions that then participate in the dialog process according to general, *domain-independent* communication rules. This is the major distinguishing aspect of the agent-based approach against the finite-state based approach. To take initiative, the dialog system just needs to import the status messages from the underlying robot system and creates an IU for it. This IU then participates in the joint common ground building process as a part of an exchange that needs to be grounded by the user. The association of such robot status messages to the communicative intentions (that are to be created for the intention layer of a new IU) can be configured externally and, thus, be easily updated or replaced. This means, we can

easily control when the robot says something and what it says. In the following subsection we describe two behaviors that are realized based on this system capability and facilitate HRI user studies.

B. The Implemented Behaviors

Before describing the implemented behaviors we want to first depict the context in which they occur. Our dialog system is implemented on our robot BIRON, a personal robot with socially 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 bought a new robot from a shop and shows it her home to prepare it for future tasks. The robot should be able to learn and remember features of objects that the user mentions and it “sees”, e.g., name, color, images etc. This is also a learning scenario for the user because she interacts with a complex robot system for the first time. Within this context we implemented two initiative taking behaviors for BIRON that can be switched on and off. With the example dialog in Fig. 1, which is transcribed from a real interaction session between a user and BIRON, we describe these behaviors and our motivation in detail.

1) *Taking greeting initiative*: In the home tour scenario the user could have difficulty to start the interaction with BIRON at the beginning. After she unpacks and turns on BIRON she would feel a little bit helpless as to how to start the interaction. We think it can help to overcome this start difficulty if BIRON can greet the user on its own. This behavior would imply for the user that the robot now works, it has already perceived the user, and it can also speak human language. Probably the user would reply to this greeting intuitively as they do in human-human interaction. (This assumption is actually confirmed in our pilot study). The implementation of this initiative is quite simple: Once the robot control system detects a human in its vicinity it sends a message to the dialog system which then initiates a Default exchange (Ex1 in in Fig. 1) to greet her (R1). Now the user is expected to provide acceptance that she heard and understood BIRON which is usually done by her reply to BIRON’s greeting (U1). In U1, the user also asks an additional question about the identity of BIRON. This question is classified as user’s initiative to create a Default Exchange that BIRON should ground by answering the question (R2). If the user had not answered BIRON’s greeting the dialog system would delete this self-initiated exchange from the discourse stack after a pre-defined time and, thus, delete the expectation that the user would reply.

2) *Making remarks on its own performance*: BIRON is a highly complex robot system including a large number of modules. Many of these modules carry out computationally expensive processing, e.g., face, gesture, and object recognition and speech recognition. The performance of many modules is also subject to environmental conditions, e.g., unexpected obstacles around can lead to failure of the follow behavior or the image processing can fail because of bad

lighting conditions. In a real user-BIRON interaction, this means, there is a variety of factors that can negatively influence the general performance of BIRON but users know nothing about. We think it can help to reduce user frustration if BIRON has the ability to show that it is also aware of its own problem and feels sorry about it. We, therefore, implemented the performance evaluation behavior for BIRON. The dialog system realizes this 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, therefore, has direct correlation to the robot's overall performance. On the other hand, the more Default exchanges there are, the better is the performance because the user and BIRON can proceed to another topic only if the current one is grounded (or deleted). Based on this performance indication BIRON makes remarks to motivate users. In our current implementation, BIRON makes remarks if it has had to initiate at least three Support exchanges for one topic or has grounded three Default exchanges in succession. In the dialog example in Fig. 1, BIRON could not understand the user's utterance twice (Ex4, Ex5) and then could not find the object specified by the user (Ex6) which resulted in the creation of three Support exchanges by the dialog system in total. According to this bad performance evaluation result the dialog system initiates the contribution R6 to motivate the user. If it were a positive evaluation result BIRON would say something like "You are really good at working with me." The wording style of these remarks is generated randomly from a set of 3 pre-defined sentences.

III. THE PILOT STUDY

In our pilot study we contrasted two verbal behavior types of BIRON in a between-subject design: *basic* and *extrovert*. The basic BIRON behaves rather passively and only says something if it is explicitly addressed by the user while the extrovert BIRON possesses *both* initiative taking abilities described above. Our goal of this pilot study was to find out if these two behavior types enabled by our dialog system can affect user impressions on our robot. If yes, this will be a strong evidence for the suitability of our system for comparative user studies on robot verbal behaviors. More specifically, the goal of this study was to test:

- 1) whether the different verbal behavior of the two types of BIRON are perceived as different;
- 2) whether the different perception of BIRON's verbal behaviors has effects on the perception of other features of BIRON such as its overall performance and interaction style;
- 3) whether the different perception of BIRON's verbal behaviors has effects on the subjects' emotional status.

A. Method

We recruited 14 subjects aged from 20 to 37 from the Bielefeld University. None of the subjects had interacted with

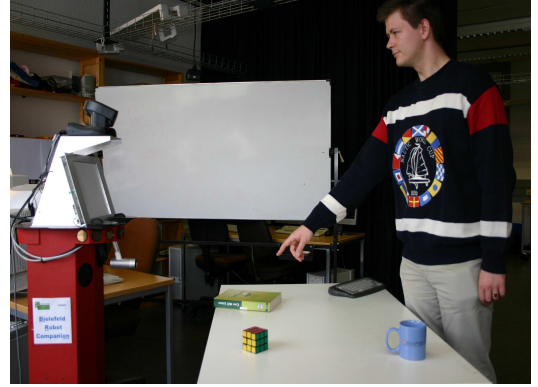


Fig. 3. The setup of the pilot study

BIRON before. In the pilot study, they were divided into two groups: 7 of them interacted with the basic BIRON (Group B) and the other 7 with the extrovert BIRON (Group E). They were just told to try out BIRON's functionalities without any knowledge about BIRON's technical limitations (Fig. 3). With this setting we intended to simulate the situation of the "innocent user" in the home tour and provoke technical failure to see users' reactions. In average, each subject interacted with BIRON more than 7 minutes. After the interaction the subjects were asked to fill out a questionnaire.

To answer the first goal question we asked subjects to rate BIRON's personality. As also assumed by [2], different behavior of a robot can cause subjects to perceive the robot as having different personality. We used 4 selected personality traits derived from the personality dimension *introversion vs. extroversion* that was proposed by Eysenck [17]. We think this dimension can be more easily associated with visible behavior and, thus, is more suitable for our purpose than his two other dimensions (*neuroticism vs. emotional stability* and *psychoticism*). The 4 traits are thoughtful vs. talkative, peaceful vs. responsive, quiet vs. active, and reserved vs. impulsive. The subjects rated BIRON's personality for each of the 4 traits using a 5-point Likert scale, e.g., 1 is very thoughtful and 5 is very talkative. For our pilot study, it is sufficient to assume that the higher value a selected item in the Likert scale has, the bigger is the tendency of the subject to classify BIRON's personality as extrovert.

To answer the second goal question we listed the four problems of our system that occur most frequently: BIRON loses the subject during the interaction, BIRON does not understand subject's utterances, the dialog with BIRON is restricted to a relatively small vocabulary and BIRON do not look in the direction of the subject's gesture. Subjects were asked to rate for each of these problems their degree of annoyance in a 5 point Likert scale. The associated questionnaire question was "How annoying were the following technical problems for you?". Additionally, we also asked users if they think the interaction style realized on BIRON is intuitive.

To measure users' emotional status after the interaction with BIRON for the last pilot study question we directly

asked them if they like BIRON.

B. Results

Given the small number of participating subjects we descriptively recorded the results of this pilot study. In the chart in Fig. 4 the result of the BIRON personality question is illustrated, the x-axis represents the personality tendency of BIRON rated by the subjects and the y-axis indicates the number of subjects who selected the corresponding items for the 4 traits. In Group B, subjects interacted with the basic BIRON and most of them thought BIRON tends to be introvert. For the Group E, the initiative taking behaviors of BIRON did lead to a clearly different picture than in Group B: extroversion is the most perceived personality tendency of BIRON even if the result is more distributed than in case of Group B.

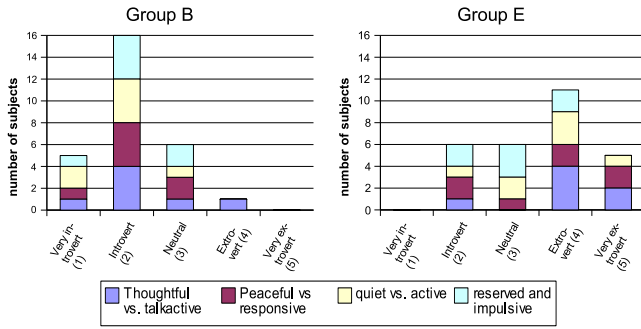


Fig. 4. The result of the question: "What do you think about the personality of BIRON?"

Figure 5 and Fig. 6 show the result of our second pilot study question. In the chart on their general annoyance degree, we can recognize a slight difference in subjects' general annoyance degree. We can even say that members of Group B seem to be generally more angry about the technical problems than those in Group E. On the issue of perceived interaction style, twice as many subjects in Group E approved the question as members of Group B.

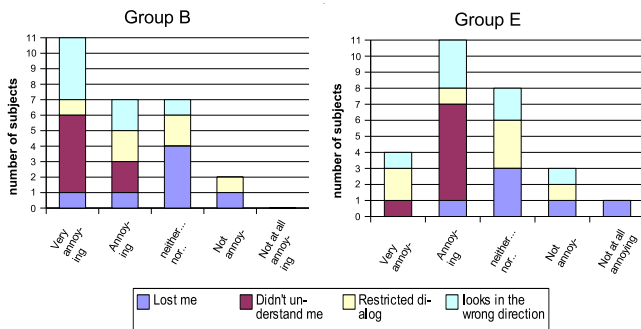


Fig. 5. The result of the question "How annoying were the following technical problems for you?"

The chart in Fig. 7 illustrates potential emotional effect of BIRON's different verbal behaviors. This result is a clear evidence for this effect: the overwhelming majority of Group

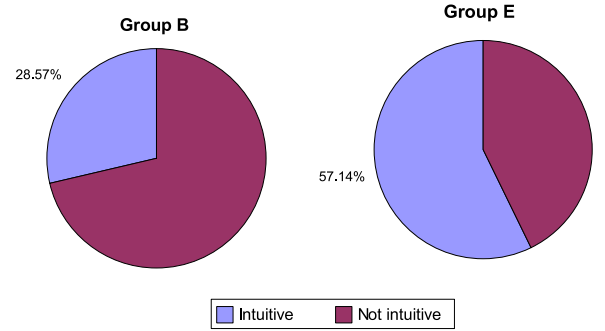


Fig. 6. The result of the question "Do you think the interaction with BIRON is intuitive?"

E liked BIRON while only a small minority of Group B held the same emotion.

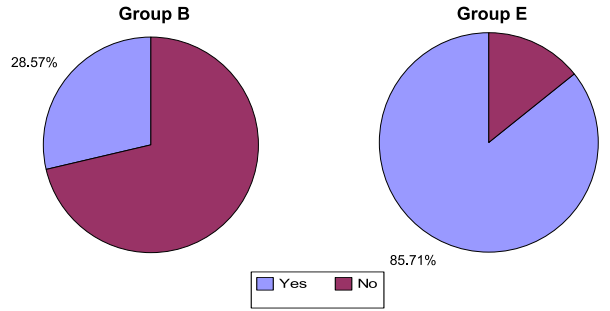


Fig. 7. The result of the question "Do you like BIRON?"

C. Discussion

The pilot study results suggest that the two verbal behavior types of BIRON were not only perceived as different, but the subjects' perception of the performance and the interaction style of BIRON were also affected, and the subjects of Group E (extrovert BIRON group) even felt emotionally animated in contrast to members of Group B. Although our study is a small-scaled pilot study, based on its clear results, we expect statistical significance of the relevant variables in larger-scaled user studies. Therefore, the pilot study results are strong evidence for the fact that our system is able to generate perceivably different verbal behaviors and is suitable for doing comparative user studies on effects of various robot verbal behaviors.

This pilot study also provides the first evidence for the benefits of a comparative user study on a real system as proposed in the section I. What we have learned includes the following two points:

- 1) *Human-awareness is important.* In the pilot study, the extrovert BIRON generally receives more positive comments from the subjects than the basic one. This result supports the advantage of human-awareness as proposed by [18] for robot motion planning domain. The extrovert BIRON demonstrates its awareness of human user's existence (by greeting her actively) and its awareness of possible emotional consequence on

users caused by its own poor performance. Such behavior does help to increase the acceptability of a robot.

- 2) *Human-awareness needs to be complemented by situation-awareness.* In our pilot study, we made the following observation: According to our experiment setup users were allowed to say anything they wanted to BIRON. But the speech understanding system of BIRON could only recognize a small number of words. Sometimes subjects were repeatedly misunderstood by the speech recognition system which caused the extrovert BIRON to initiate apology every 3 utterances. This behavior was in general rated as good by the subjects, but in this specific situation, the subjects were already annoyed by the speech recognition problem and, therefore, considered this kind of initiative as bothering. This example shows that a robot does not only need to be aware of human but also be aware of the situation and should make decisions taking into account observed situational cues.

To summarize the benefits of the comparative user study on a real system: if our pilot study had not been a comparative study, we might be able to assume the general benefits of human-awareness (as described in the first point), but we would not be able to quantify this difference to this extent which greatly helps to intensify our insight. Equally, if our pilot study had not been conducted on a real robot system with its inherent technical instability, we might not even be able to realize the complementary importance of situation awareness. Thus, our pilot study serves as a good example for the practical benefits of combining WoZ comparative studies with real robot testing.

IV. CONCLUSION

In this paper we presented a dialog system which can be easily configured to demonstrate various initiative taking behavior. In the pilot study conducted with this system the enabled verbal behaviors of our robot BIRON were clearly perceived as different and these differences also affect other factors of the user-BIRON interaction. Our system, thus, provides a suitable platform for doing comparative user studies on robot verbal behavior. Additionally, the combination of the advantages of comparative WoZ study with real robot testing also turned out to be productive in terms of gaining general insights in HRI.

To be used in comparative user studies on robot verbal behavior in general, our system is not confined to the implemented behaviors as described in this paper. It will be quite simple to enable BIRON to take initiative of other types, e.g., a task planner can provide task specific information to the dialog system so that the robot can take task-related initiative, this means, it can potentially make constructive suggestions when doing tasks collaboratively with the user. Besides, initiative behavior can generally be parameterized to adapt to the robot's knowledge of the current social situation, e.g., the dialog system can be configured in a way that it only takes greeting initiative when the robot's vision system can recognize user's willingness to interact with the robot. Since

the initiative taking behavior can be easily modified in our system, we can, as well, contrast verbal behaviors based on the principles above in comparative user studies.

V. ACKNOWLEDGMENTS

This work is funded by the European European Commission Division FP6-IST Future and Emerging Technologies within the Integrated Project COGNIRON (The Cognitive Robot Companion) under Contract FP6-002020.

REFERENCES

- [1] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behaviour to tasks to improve human-robot cooperation," in *Proc. Int Symposium on Robot and Human Interactive Communication*, 2003.
- [2] S. Woods, K. Dautenhahn, C. Kaouri, R. te Boekhorst, and K. L. Koay, "Is this robot like me? links between human and robot personality traits," in *Proc. IEEE-RAS International Conference on Humanoid Robots*, 2005.
- [3] M. L. Walters, K. Dautenhahn, K. L. Koay, C. Kaouri, R. te Boekhorst, C. L. Nehaniv, I. Werry, and D. Lee, "Close encounters: Spatial distances between people and a robot of mechanistic appearance," in *Proc. Int. Conf. on Humanoid Robots*, 2005.
- [4] K. Severinson-Eklundh, A. Green, and H. Htenrauch, "Social and collaborative aspects of interaction with a service robot," *Robotics and Autonomous Systems*, vol. 42, 2003.
- [5] T. Kanda, T. Hirano, and D. Eaton, "Interactive robots as social partners and peer tutors for children: A field trial," *Human-Computer Interaction*, vol. 19, 2004.
- [6] S. Li, M. Kleinhagenbrock, J. Fritsch, B. Wrede, and G. Sagerer, "'BIRON, let me show you something': Evaluating the interaction with a robot companion," in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, Special Session on Human-Robot Interaction*, 2004.
- [7] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu, "Integrated natural spoken dialogue system of jijo-2 mobile robot for office services," 1999.
- [8] R. Bischoff and V. Graefe, "Dependable multimodal communication and interaction with robotic assistants," in *Proc. Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, 2002.
- [9] I. Toptsis, S. Li, B. Wrede, and G. A. Fink, "A multi-modal dialog system for a mobile robot," in *Proc. Int. Conf. on Spoken Language Processing*, 2004.
- [10] K. Aoyama and H. Shimomura, "Real world speech interaction with a humanoid robot on a layered robot behavior control architecture," in *Proc. Int. Conf. on Robotics and Automation*, 2005.
- [11] M. F. McTear, "Spoken dialogue technology: enabling the conversational interface," *ACM Computing Surveys*, vol. 34, no. 1, 2002.
- [12] H. H. Clark, Ed., *Arenas of Language Use*. University of Chicago Press, 1992.
- [13] D. Traum, "A computational theory of grounding in natural language conversation," Ph.D. dissertation, University of Rochester, 1994.
- [14] 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.
- [15] 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*, 2005.
- [16] S. Li, B. Wrede, and G. Sagerer, "A computational model of multimodal grounding for human-robot interaction," in *Proc. Int. SIGdial Workshop on Discourse and Dialogue*. ACL, 2006.
- [17] H. J. E. . S. B. G. Eysenck, *Manual of the Eysenck personality questionnaire*. Hodder and Stoughton, 1975.
- [18] K. Dautenhahn, S. Woods, M. Walters, K.-L. Koay, C. Nehaniv, E. Sisbot, R. Alami, and T. Simeon, "How may i serve you? a robot companion approaching a seated person in a helping context," in *Proc. int. Conf. on Human Robot Interaction*, 2006.

BIRON, what's the topic?

A Multi-Modal Topic Tracker for improved Human-Robot Interaction

Jan Frederik Maas, Thorsten Spexard, Jannik Fritsch[†], Britta Wrede and Gerhard Sagerer

Faculty of Technology, Applied Computer Science

Bielefeld University

33594 Bielefeld, Germany

Abstract—Creating robots with extendable social skills and interaction capabilities that suffice their operation in the real world with naive users is a very challenging task. In this paper we present a new approach using topic tracking on multi-modal dialogue to provide a mobile robot with a higher level situation awareness in human-robot interaction. The robot is no longer operating in laboratory surroundings, but in its own real world flat. We describe how our topic tracking approach is implemented in this integrated system, operating on verbal speech input. Different modalities like data from video cameras and laser scans are used as additional cues to a semantic understanding and grouping of user utterances into different topics. Both the amount of topics and the according topic names are created dynamically. Evaluating an offline speech corpus demonstrates the suitability of our approach. It is now possible to ask “BIRON, what's the topic?”, making the interaction more social.

I. INTRODUCTION

What is a social robot and how is a real robot companion discriminated against a more common service robot performing fetch and carry tasks only? This question is still in discussion [1]. Mobile robots that participate in our daily life have to satisfy many requirements. Mainly they have to find and track humans, operate in various surroundings with robust localisation and navigation and therefore have to adapt or learn how to cope with unknown situations.

One first step towards this goal is the so called Home Tour Scenario, developed within the *COGNIRON* (Cognitive Robot Companion) project [2]. In this scenario a person with unspecified knowledge of robotics has just bought a robot companion and gives it a tour through her private home so to familiarise it with its new habitat. The human points to and names locations and objects which she believes are necessary for the robot to remember.

Our mobile demonstrator *BIRON* (Bielefeld Robot Companion) [3] successfully accomplishes this task in a real flat. But robots moving out of laboratories have to gain more social skills to improve human-robot interaction. Besides the verbal information, in human-human interaction many additional cues such as gestures, mimics, and context knowledge are used intuitively to achieve a successful communication. Even if the mere verbal information is underspecified or

inconsistent in most cases both interaction partners know which topic they are talking about.

Using a natural spoken dialogue system, it is crucial for the acceptance of a robot by humans that it understands the user's intention. The robot has to react comprehensible even if the information given by user utterances was imprecise. According to the human example, knowledge about the recent conversation topic helps to resolve underspecified real world references. For example, the user may ask the robot to find a brush, but does not specify if she wants a brush for cleaning the dishes or making her hair. Depending of the topic the recent conversation is about, the robot can decide by itself which brush to look for or suggest which brush it will choose.

For the evolution of sociable robots, knowing what a conversation is about is also crucial for research areas like *situation awareness* [4]. Robots will not only have to reflect about basic sensory data as in many reactive systems where situation awareness is mainly used for tasks like navigation in the presence of humans or object tracking. Situation awareness with regard to a more complex human-robot interaction and verbal communication is a basis for successful joint or shared attention abilities [5], [6]. Based on the knowledge what a person is often talking about robots can learn the user's preferences, e.g., taking the initiative and telling the user the recent sports results, if the user previously often talked about sports. By this means robots will develop from pure task fulfilling machines towards companions.

In this paper it is described how we realized a Dynamic Topic Tracking (DTT) of naturally spoken conversation between human and robot. Though user utterances only would suffice for the proper operation of the described module, we use different additional cues derived from the current system state and audio-visual input to enhance the reliability in real world settings. Situation awareness is realized not only using classical sensors like video cameras and laser range finders, but additionally interpreting natural speech in a human-robot dialogue.

The paper is organised as follows. In section II we will give an overview of work related to topic detection and situation awareness especially in the context of mobile robots. Subsequently it is described how the DTT is linked to the overall system architecture in section III, followed by

[†]J. Fritsch is now with the Honda Research Institute Europe GmbH in Offenbach, Germany.

a short description of the modules delivering data for the DTT in section IV. The DTT module itself is presented in detail in section V and results of an evaluation are given in section VI. The paper ends with a conclusion in section VII summarising the results and presenting an outlook on future work.

II. RELATED WORK

Topic tracking on human-robot communication during situated communication is a new area of research, which has to meet special requirements: For example, the topic tracking has to be *online*, i.e., the results of the tracking must be delivered in real time. Additionally, the high dynamic of household environments demands for a *dynamic* approach - the topics have to be learned during communication, and the algorithm should not rely on pre-defined or pre-trained topics.

In the past, much topic tracking research has been carried out in the context of Information Retrieval [7]. Information Retrieval is centred around making large corpora – such as newspaper corpora – accessible for queries, e.g., automatically finding all news belonging to a certain topic. While the background is very different, standard techniques such as lemmatisation can be applied to topic tracking on a robot, too.

A related area of research is Text Segmentation, i.e., automatically structuring unsegmented text into segments which express exactly one topic, while subsequent segments bear different topics. Since topic tracking in human-robot interaction also deals with topically unsegmented streams of utterances, we use algorithms (such as Latent Semantic Analysis [8]) which proved to be highly efficient for this task [9].

General approaches for dynamic topic tracking on spoken language, i.e., broadcast news corpora, have been developed in the context of the Topic Detection and Tracking project [10] amongst others, for example [11]. The task of segmenting spoken language into topics by using different cues has also been researched [12], [13].

To achieve sophisticated human robot interaction, besides information retrieval, topic tracking delivers important data for the intention of a human and the situation a robot is involved in. Work has already been carried out in determining the intention of a human by classifying gestures in different categories like symbolic or referential gestures [14].

In [15] situation awareness is described in the context of a mobile robot for urban search and rescue tasks. Depending on the situation it finds itself in, the robot has to separate important information from less important one so to provide only important data to the operator for a quick decision taking process. Although the system is able to classify a situation this process is based on the system state and information retrieved from auditive and visual sensors only. Instead of a higher level user interaction an unidirectional command execution is performed.

Taking the reactions of a user more into account, [16] presents an approach in which the robot is aware of the

level of engagement a human has towards a robot during an interaction. This is to ascertain that a human does not leak of attention while the robot performs different subtasks. For the evaluation physiological signals like cardiovascular, electro dermal, and electromyogram activity of the interaction partner have to be gathered by various sensors connected to the user which is far away from human-like communication.

Investigating topic tracking on user utterances with respect to integrated robot systems it seems that there is still little scientific work done in combining these aspects to achieve a more high level situation awareness. Furthermore situation awareness in human robot interaction is still very often connected not with the robot but with the human getting more information by the robot. We propose that the ability of a robot to understand what an interaction is about and to accordingly react, enhances its sociability. How a topic tracking is integrated into the current architecture will be described in the following sections.

III. ARCHITECTURE

For the presented system a three layer architecture [17] is used consisting of a deliberative, an intermediate, and a reactive layer (s. Fig. 1). The DTT is located in the top deliberative layers well as a dialog system for complex user interaction. In the bottom layer reactive modules capable of adapting to sudden changes in the environment are placed. Since neither the deliberative layer dominates the reactive layer nor the reactive layer dominates the deliberative one, a module called Execution Supervisor (ESV) was developed [18] located in the intermediate layer, where the Localisation and its knowledge base are also positioned. The ESV coordinates the different tasks of the individual modules by reconfiguring the parameters of each module. For example, the Player Interface for controlling the hardware is configured to receive movement commands from different modules.

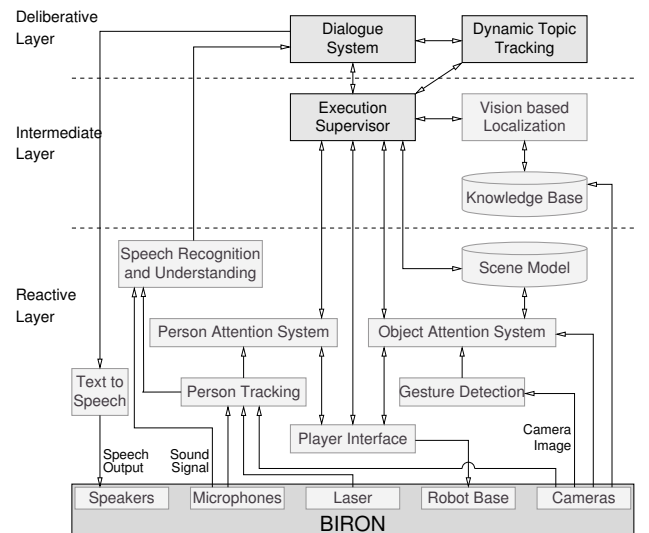


Fig. 1. Integration of DTT into the HRI architecture of BIRON.

The ESV can be described as a finite state machine. The different HRI abilities are represented as states and a message sent from a module to the ESV can result in a transition from state A to state B. For each transition the modules in the different layers are reconfigured. Additionally to the new configuration, data like movement speed is exchanged between the modules. All data exchange via the ESV is based on the XML Communication Framework (XCF, see [19]) using four predefined XML structures, only. All XML data is designed in a human readable style for easy understanding of the internal system communication and efficient debugging.

Using a predefined set of limited XML structures data exchange between the ESV and each module is automatically established after reading a configuration file. The file also contains the definition of the finite state machine and the transitions that can be performed. This makes the system easily extendable for new HRI capabilities, by simply changing the configuration file for adding new components like the DTT without changing one line of source code. Due to the automatic creation of the XML interfaces with a very homogenous structure, fusing the data from the different modules is achieved easily.

The system already contained modules for multiple person tracking with attention control [20],[21] and an object attention system [22] based on deictic gestures for learning new objects. Additionally a vision based localisation [23], was added to teach the robot names of different rooms and asking it for its recent position.

Here we detail how the existing human-robot interaction is improved by using topic tracking, analysing not only system state and visual cues, but spoken language via a dialog system [24] as well. To achieve this functionality, the integrated system is extended and the DTT module is added to the overall architecture as depicted in Fig. 1.

Besides the association of pure verbal input to different topics for a dynamic creation of different topics the flexible design of the architecture enables the DTT to use additional information provided by the various modules to enhance the topic tracking. A more detailed description of these modules and their data provided for the DTT is given in the following.

IV. MODULES FOR DTT

The most significant advantage for an integrated topic tracking system on a mobile robot is the possibility to gather multi-modal information from the different sources available on a robot system. For our topic tracking system, we primarily use this information to perform two different tasks (For a more detailed discussion, see chapter V below.):

- **Segmentation:** The system needs to indicate topic changes, i.e., it has to learn which consecutive user utterances belong to the same topic.
- **Tracking:** The system needs – of course – data that it can classify as belonging to a topic. For our system, these are speech data (i.e., words) and object references (verbal or gestural) from the user. It is possible to enlarge this set, though (see chapter VII).

To gather this information, the topic tracking system receives data from the following modules:

A. Speech Recognition, Speech Understanding and Dialogue (Tracking and Segmentation)

The speech recognition system translates the user utterances into words, which can be directly tracked by the topic detection system. The speech understanding allows in combination with the dialogue system to directly treat explicit topic changes and references (e.g., “let us talk about plants”), as well as implicit topic changes (e.g., “follow me” often indicates a topic change). Consider that although explicit topic references are possible, they are not necessary for our system.

Additionally, discourse structure information from the dialogue system is used to topically segment the user utterances.

B. Gesture- and object recognition (Tracking)

In combination with the dialogue system, the gesture- and object recognition modules deliver information about objects referred to by gestures and verbal description of the user. Object references (e.g., IDs) can – as well as words – be recognised as belonging to a certain topic, therefore delivering additional cues for topic tracking.

C. Person Attention System (Segmentation)

The person attention system (PTA) tracks multiple humans by fusing information from different sensors. Laser scans are used for classifying pairs of legs, a speaker localisation is performed on audio data, and both position and gazing direction of faces are detected on video images. Subsequently the person most likely to interact with the robot is selected and data concerning this person is delivered. Especially information about the user facing the robot or not is relevant, because phases of looking away often indicate reorientation of the user, after which commonly¹ a topic shift occurs.

D. Execution Supervisor (Segmentation)

The ESV indicates whether a user is in direct communication with the robot or not. Phases of no direct communication clearly indicate topic “shifts”. Additionally, the ESV indicates phases of object interaction, during which the user looks away (i.e., on the object) from the robot. These phases usually do not result in topic shifts, therefore the ESV information can correct the segmentation by the PTA.

The cues for segmentation of user utterances into thematic blocks are heuristically evaluated by the topic detection system, while the tracking data is directly processed. After elaborating on the information sources of our topic tracking system, we will now describe its structure and functionality.

¹at least in a home tour scenario

V. DYNAMIC TOPIC TRACKING

As mentioned above, a topic tracking algorithm for situated human-robot communication has to meet several requirements: It has to work in real time (online), the topics should not be predefined but learned during the communication (dynamic) and the algorithm must be able to incorporate the different information sources a robotic system can access, otherwise it would waste valuable information.

The main information source for a topic tracking system is spoken language, though. The utterances of the user are to be classified as belonging to a known or unknown topic, based on the information the robot learned during the current or past communications.

Usually – i.e., in the context of Information Retrieval or Text Segmentation tasks – the main cues which indicate the topic of an utterance are words. For example, the word “soccer” strongly indicates a sport topic in a newspaper. A great advantage of our approach is that not only words can be processed, but other symbols, such as IDs of objects referred to, as well [25]. Therefore, multi-modal topic cues can be considered in our system.

To give an example: Because of the ability of our robot system to recognise objects denoted by the user, information about object identity can therefore be considered by the topic tracking system. This way, it can differentiate between a knife used in the kitchen and a knife used in the garden, both belonging to different thematic areas.

In our topic detection system, the basis for deciding which symbol (word, object ID) belongs to which topic are symbol co-occurrences – symbols which co-occur often in the same contexts (i.e., segments, see paragraph about segmentation below) can be considered as bearing the same topic, while symbols seldom co-occurring probably belong to different topics. This way, topics can be built dynamically by clustering symbols into sets (see “clustering” below).

Roughly following the approach of the Topic Detection and Tracking project [10], the main tasks of topic detection systems are: Preprocessing, Segmentation, Semantic Association, Clustering and Tracking. We will elaborate on these aspects in the following paragraphs.

Preprocessing Three main tasks have to be accomplished during preprocessing: First, words indicating no topic, i.e., function words, should be deleted. This is done by using a stoplist of words not to be processed. Second, words should be reduced to their stems or lemmas, thus merging words bearing the same topic. This task is done by a simple lookup, although much more elaborated lemmatisation algorithms are available. For example, for the offline evaluation, the Tree Tagger [26] was used for lemmatisation. While these two steps are standard techniques for Information Retrieval tasks, the third is somewhat unique for our system and only possible because of the situated information available: Words expressing different topics should be split up, which is done by substituting verbal references to objects by the IDs of the object. Consider that this way, words referring to a single object are disambiguated, thus additionally supporting the

second preprocessing task, i.e., merging words bearing the same topic.

Segmentation It is necessary to split the stream of user utterances into segments, which ideally bear a single topic. Information about topic changes can be utilised for the tracking process, but is also the basis for training the topic models, because words co-occurring in many segments can be assumed to bear the same topic. In our system, segmentation is done heuristically and based on multi-modal cues, as mentioned in chapter IV.

Semantic Association From the co-occurrence data a so called Semantic Space is build. Semantic spaces are vector spaces in which a vector represents a single word (respectively object ID)². The distance between the vectors indicates the similarity of the words or – in our case – the thematic distance.

For our system, we implemented three different Semantic Spaces: The simple vector space model [27], Fuzzy Semantics [28] and Latent Semantic Analysis (LSA) [8]. We did not implement the popular Probabilistic Latent Semantic Analysis (PLSA), because for this method the number of topics must be specified beforehand.

Clustering By using distance information, the semantic spaces are clustered in thematic areas which ideally bear a single topic. For our purposes we use complete-linkage agglomerative clustering, which has proved to be fast, delivering good results [29]. As distance measures we implemented cosine distance, correlation and euclidian distance, although numerous other are possible. We usually apply correlation, because it distinguishes between correlated and uncorrelated elements.

Tracking The information about topic clusters can be used to estimate the topic for each user utterance. For each word of an utterance and for each topic a distance from the word to the centroid of the topic is computed. We abstained from calculating normal distributions, because in most cases not enough training material is available to robustly calculate the parameters. The topic with the lowest sum of added distances and history information is the detected topic.

Additionally, if no topic change is detected between the current and the last user utterance, the system is biased towards the last detected topic, thus implementing a history. This is done by multiplying the sums of added distances of the last utterance with 0.3 and adding them to the current distances. Consider that if an utterance contains no topic indicating symbols, the last topic is automatically indicated this way.

Preprocessing, segmentation and tracking are performed in real time. The computation of the Semantic Spaces and the clustering process can take too much time for real time calculations, though. For this reason, these steps are continually computed in the background, updating the tracker when ready. This update process usually takes no more than a few seconds for normal-sized dialogues.

²In other research contexts, Semantic Spaces are also applied to represent documents



Fig. 2. Human-robot interaction in the BITT environment.

VI. EVALUATION

It is our goal to show the feasibility of our topic tracking system for communication sequences of different length. For this reason, an offline evaluation on a corpus of human robot interaction was performed, as described in the next section. Online experiments of short human robot communication sequences are carried out at the moment, and are detailed out subsequently.

A. OFFLINE

In order to estimate the efficiency of the topic detection system in the home tour scenario, the topic tracking was evaluated on the BITT-corpus [30].

Corpus The corpus consists of medium-sized human-robot interaction sequences of 29 subjects describing a room to the robot BIRON. The room contains several thematic areas the subjects could elaborate on, thus resembling a household environment. An impression of an interaction between a human and BIRON is given in Fig. 2. In the background, a working place with a computer and a small kitchenette are visible, both being thematic areas for the experiments.

Although BIRON was enabled to track the subjects, thus simulating attention and performing as a communication target, we decided to deactivate the dialogue capabilities of the robot in order to not constrain the subjects utterances by a constrained dialogue. Thus, the corpus contains monologues of free speech of the subjects. Consider that free speech puts additional workload on a topic tracking system compared to constrained speech with a dialogue system.

The corpus contains manually transcribed speech in which references to objects were annotated. Lemmatisation was performed with the Tree Tagger tagging system [26]. This way, the corpus delivers the result of the preprocessing steps of the topic tracking except for deletion of function words, which was carried out by a stoplist.

The corpus consists of user utterances, which were detected by the same automatic voice activity detection we use on BIRON and manually transcribed afterwards. For evaluation purposes, three annotators assigned one or no topic to each of these utterances. The annotators were given no predefined topics, although they had to determine a set of topics before carrying out the annotation (and after making themselves familiar with the corpus). This way, the task of the annotators strongly resembled the task of the topic tracking system, i.e. freely building topics and assigning them to utterances. On average, about 4900 utterances were assigned an topic by each annotator with about seven to eight topics mean per dialogue. No annotator used more than ten different topics for the whole corpus. Usually, a manually annotated topic occurred only one or two times during a monologue.

Evaluation process For the purpose of evaluation, the manually annotated topic information was also utilised for simulating the segmentation process, i.e., two consecutive utterances bearing the same topic were considered to belong to the same segment.

During the offline evaluation, the history – as described in chapter V – was activated. Consider that for the offline evaluation, the history was only deleted between – and not during – processing monologues. Thus, the topic change information gained from the topic annotation was only used for training but not for tracking purposes.

In our experiments we distinguished between a word-based approach and a mixed one. For the word-based approach, only lemmatised words uttered by the subjects were used for training and tracking. For the mixed approach, object references by the subjects were substituted by object IDs, thus simulating the capability of our robot to detect and recognise objects. This way, we can estimate the gain by utilising situated information.

During performing evaluation for the mixed approach, words were given a 50% penalty during tracking in comparison to object IDs, because object IDs proved to be more reliable cues for indicating topics.

The topic detection system tracked each of the 29 monologues after being trained with the co-occurrence data built from the other monologues (bootstrapping). The clustering algorithm was biased to generate approximately as many topics as the annotators. This was accomplished by setting the maximum amount of clusters to be detected to ten, slightly more than the maximum amount of manually detected clusters.

After tracking, the results were compared to each of the three manual annotations. As an evaluation measure, f-score [31] was utilised. f is gained by averaging precision and recall as shown in (1):

$$f = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \quad (1)$$

Consider that the topic tracking algorithm was forced to detect a topic whenever possible. Consequently, utterances

TABLE I
AVERAGED F – WORDS ONLY

<i>C</i>	<i>Vector Space</i>	<i>Fuzzy Semantics</i>	<i>LSA</i>
2	0.4	0.46	0.52
5	0.53	0.48	0.54
10	0.5	0.51	0.51
20	0.41	0.41	0.44

TABLE II
AVERAGED F – MIXED APPROACH

<i>C</i>	<i>Vector Space</i>	<i>Fuzzy Semantics</i>	<i>LSA</i>
2	0.51	0.55	0.67
5	0.7	0.49	0.69
10	0.7	0.56	0.71
20	0.51	0.54	0.52

which were not manually noted as bearing a topic were not taken into account for calculating precision.

For the calculation of precision and recall it was necessary to determine for each manually annotated topic the corresponding automatically detected topic which is considered to be correct. As in [32], the correct automatically detected topic for a manually annotated topic was determined by comparing the tracking results with the manual annotation. The automatically detected topic which covered the highest amount of utterances of a manually annotated topic was considered to be correct. All utterances of the manually annotated topic which were indicated as bearing another automatically tracked topic were consequently considered to be falsely tracked. Afterwards, the f-values for each monologue, human annotation and manually annotated topic were simply averaged by calculating the arithmetic mean.

To compute inter-annotator agreement, for each of the six pairs of human annotations the same evaluation metric was applied. The averaged f for these computations was 0.79, indicating a medium agreement between the human annotators.

For the different Semantic Spaces, the results presented in Table I and Table II were achieved. Table I shows the results for word-based tracking only. Table II shows the results for the mixed approach in which verbal object references were replaced by object IDs. As mentioned, each cell shows the arithmetic mean of the f-values scored for each of the three manual annotations, monologue and manually annotated topic. *C* is the minimum number of contexts a object reference or a word had to occur in to be considered during the calculation. Increasing *C* results in less noisy training data, since seldom occurring symbols are deleted. It results in loss of information as well, since the topic affiliation of these symbols is lost for the processes utilising topic information. Additionally, very high values of *C* (i.e., $C > 15$) will result in reduced tracking results since valuable cues (symbols) for topic tracking are not considered.

Discussion As expected, LSA performed best in the topic tracking task with a maximum f of about 0.71 and best performance with a low *C*. This can be explained with the enhanced capability of LSA to deal with noisy data compared

to a simple vector space.

The unmodified vector space shows promising results, too, especially with medium values for *C*. Since the results for the simple vector space and LSA converge for higher values of *C*, the noise in the data seems to consist mainly of words and object references with a low frequency.

The Fuzzy Semantics Space did not deliver satisfying results in comparison. This may be explainable by a conflict between the clustering algorithm approach and this semantic space.

Our results show clearly that utilising situated information in the form of object references strongly enhances the topic tracking outcome. Additionally, for the purpose of resolving underspecified object references – as mentioned in the introduction – the computation of topics for object references is needed. Thus, the mixed approach is to be considered superior in comparison to a simple word-based approach.

B. ONLINE

The integration of the topic tracking system on the robot BIRON enables us to perform topic tracking online, i.e., during communication. The detected topic can directly be queried, e.g., by asking “BIRON, what’s the topic?”.

At the moment, experiments are carried out. The experiments will focus on the performance of the topic tracking system in short, situated communications of unexperienced users with our robot system. A new feature of our system will be tested: The users have the possibility to make topics and topic changes explicit, i.e., “let us talk about the kitchen”. Although explicit topic changes were part of the BITT corpus, they were not taken into account during the offline evaluation. Adding explicit topics and topic changes will lead to improved communication capabilities of our robot system.

Although the experiments are not finished, we could already prove that subjects communicating with BIRON about several sets of thematically connected objects accredited BIRON the capability of learning discourse coherence. The experiments show a significant (<5%) difference between subjects communicating with BIRON with and without an activated topic tracking system.

VII. CONCLUSION

The offline experiments demonstrate that our approach is feasible for communication sequences of medium length. The results of the experiments encourage us to carry out further studies, as described in the former paragraph.

Additionally, the performance of our topic tracking system can be enhanced by taking into account additional contextual information, such as information about the current location, which can be attained by BIRON’s vision based localisation system. As mentioned, [25] has shown that additional features from different modalities can enhance the performance of an LSA and probably for a vector space as well.

Further, our topic tracking system makes it possible to approach the resolution of underspecified object references, e.g., “find the brush”, when the type of the brush is not clearly specified. This can directly be accomplished by the

dialogue system querying the topic tracking, thus getting information about the current topic. Although it is already possible to query the current topic by asking “BIRON, what’s the topic?”, the interaction with the robot will really benefit when other modules make more extended use of the topic tracking results.

ACKNOWLEDGEMENTS

This work has been supported by the European Union within the ‘Cognitive Robot Companion’ (COGNIRON) project (FP6-IST-002020) [2] and by the German Research Foundation within the Graduate Program “Task Oriented Communication”.

REFERENCES

- [1] K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. Koay, and I. Werry, “What is a robot companion - friend, assistant or butler?” in *Proc. IEEE IRS/RSJ Int. Conference on Intelligent Robots and Systems*, Edmonton, Alberta, Canada, 2005, pp. 1488–1493.
- [2] COGNIRON, “The cognitive robot companion,” <http://www.cogniron.org>.
- [3] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Tóptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, “BIRON – The Bielefeld Robot Companion,” in *Proc. Int. Workshop on Advances in Service Robotics*, E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, Eds. Stuttgart, Germany: Fraunhofer IRB Verlag, May 2004, pp. 27–32.
- [4] J. Crowley, “Context driven observation of human activity,” in *Proc. of the European Symposium on Ambient Intelligence*, Amsterdam, The Netherlands, 2003, pp. 101–118.
- [5] C. Breazeal, *Designing Sociable Robots*. Cambridge, MA: MIT Press, 2002.
- [6] Y. Nagai, “Understanding the development of joint attention from a viewpoint of cognitive developmental robotics,” Ph.D. dissertation, Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, 2004.
- [7] C. van Rijsbergen, *Information Retrieval*, 2nd ed. Internet Address, 2005. [Online]. Available: <http://www.dcs.gla.ac.uk/Keith/Preface.html>
- [8] T. K. Landauer, P. W. Foltz, and D. Laham, “Introduction to latent semantic analysis,” *Discourse Processes*, vol. 25, pp. 259–284, 1998.
- [9] P. W.-H. Freddy Choi and J. Moore, “Latent semantic analysis for text segmentation,” in *Proceedings of 6th EMNLP*, 2001, pp. 109–117.
- [10] J. Allan, Ed., *Topic Detection and Tracking*. Norwell, Massachusetts: Kluwer Academic Publishers, 2002.
- [11] S. Renals, D. Abberley, D. Kirby, and T. Robinson, “Indexing and retrieval of broadcast news,” *Speech Communication*, September 2000.
- [12] J. Ponte and W. Croft, “Text segmentation by topic,” in *ECDL ’97: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*. Morgan Kaufman Publishers, 1997, pp. 113–125.
- [13] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, “Prosody-based automatic segmentation of speech into sentences and topics,” *Speech Communication*, vol. 1–2, no. 32, pp. 127–154, 2000.
- [14] C. L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Hägele, C. Parlitz, and R. Alami, “A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction,” in *Proc. IEEE Int. Symp. on Robot and Human Interactive Communication*, 2005, pp. 371–377.
- [15] A. Carbone, A. Finzi, A. Orlandini, F. Pirri, and G. Ugazio, “Augmenting situation awareness via model-based control in rescue robots,” in *Proc. IEEE Int. Conf. on Intelligent Robots and Systems*, 2005, pp. 1549–1555.
- [16] P. Rani and N. Sarkar, “Operator engagement detection and robot behavior adaption in human-robot interaction,” in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2005, pp. 2063–2068.
- [17] J. Fritsch, M. Kleinehagenbrock, A. Haasch, S. Wrede, and G. Sagerer, “A flexible infrastructure for the development of a robot companion with extensible HRI-capabilities,” in *Proc. IEEE Int. Conf. on Robotics and Automation*, Barcelona, Spain, April 2005, pp. 3419–3425.
- [18] M. Kleinehagenbrock, J. Fritsch, and G. Sagerer, “Supporting advanced interaction capabilities on a mobile robot with a flexible control system,” in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, vol. 3, Sendai, Japan, September/October 2004, pp. 3649–3655.
- [19] S. Wrede, J. Fritsch, C. Bauckhage, and G. Sagerer, “An XML Based Framework for Cognitive Vision Architectures,” in *Proc. Int. Conf. on Pattern Recognition*, no. 1, 2004, pp. 757–760.
- [20] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, “Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot,” in *Proc. Int. Conf. on Multimodal Interfaces*. Vancouver, Canada: ACM, November 2003, pp. 28–35.
- [21] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer, “Audiovisual person tracking with a mobile robot,” in *Proc. Int. Conf. on Intelligent Autonomous Systems*, F. Groen, N. Amato, A. Bonarini, E. Yoshida, and B. Kröse, Eds. Amsterdam: IOS Press, March 2004, pp. 898–906.
- [22] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer, “A multi-modal object attention system for a mobile robot,” in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. Edmonton, Alberta, Canada: IEEE, August 2005, pp. 1499–1504.
- [23] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Kröse, “Biron, where are you? - enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization,” in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, October 2006, to appear.
- [24] S. Li, B. Wrede, and G. Sagerer, “A computational model of multi-modal grounding,” in *Proc. SIGdial workshop on discourse and dialog*. ACL Press, 2006.
- [25] R. Serafin and B. D. Eugenio, “Flsa: Extending latent semantic analysis with features for dialogue act classification,” in *ACL 2004*, 2004, pp. 692–699.
- [26] H. Schmid, “Probabilistic part-of-speech tagging using decision trees,” in *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.
- [27] G. Salton, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [28] B. B. Rieger, “Feasible fuzzy semantics. on some problems of how to handle word meaning empirically,” in *Words, Worlds and Contexts*, H. R. H.J. Eikmeyer, Ed. de Gruyter, 1981, pp. 193–209.
- [29] P. Cimiano, A. Hotho, and S. Staab, “Comparing conceptual, partitioned and agglomerative clustering for learning taxonomies from text,” in *Proceedings of the European Conference on Artificial Intelligence (ECAI’04)*. IOS Press, 2004, pp. 435–439.
- [30] J. F. Maas and B. Wrede, “BITT: A corpus for topic tracking evaluation on multimodal human-robot-interaction,” in *Proceedings of the international conference on Language and Evaluation (LREC)*, Genoa, Italy, 2006, to appear.
- [31] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining,” 2005, <http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf>.
- [32] Y. W. Seo and K. Sycara, “Text clustering for topic detection,” Robotics Institute, Carnegie Mellon University, Tech. Rep., January 2004.

BITT: A Corpus for Topic Tracking Evaluation on Multimodal Human-Robot-Interaction

Jan Frederik Maas*, Britta Wrede*

*Bielefeld University
Technical Faculty, Applied Computer Science Group
Universitätsstr. 25
33615 Bielefeld
{jmaas, bwrede}@techfak.uni-bielefeld.de

Abstract

Our research is concerned with the development of robotic systems which can support people in household environments, such as taking care of elderly people. A central goal of our research consists in creating robot systems which are able to learn and communicate about a given environment without the need of a specially trained user. For the communication with such users it is necessary that the robot is able to communicate multimodally, which especially includes the ability to communicate in natural language. We believe that the ability to communicate naturally in multimodal communication must be supported by the ability to access contextual information, with topical knowledge being an important aspect of this knowledge. Therefore, we currently develop a topic tracking system for situated human-robot communication on our robot systems. This paper describes the BITT (Bielefeld Topic Tracking) corpus which we built in order to develop and evaluate our system. The corpus consists of human-robot communication sequences about a home-like environment, delivering access to the information sources a multimodal topic tracking system requires.

1. Introduction

Our research takes place within the “Cognitive Robot Companion” (COGNIRON¹) project. The project is concerned with the development of robotic systems which can support people in household environments, for example by taking care of elderly people, or performing everyday tasks.

A central milestone of the project is to create robot systems which are able to learn and communicate about a given environment without the need of a specially trained user. Thus, our research is focused on building a robotic system being capable of multimodal communication, especially natural language. The communication capabilities of the robot system should be as natural as possible, because constrained communicating systems – such as dialogue systems not capable of proper anaphora resolution – put additional workload on a non-specialist communication partner. Based on these considerations, we designate human-robot interaction as “natural” when

1. the used modalities (language, gestures, etc.) are the same as for face-to-face human-human communication
2. the human does not have to learn how to communicate with the robot, but can apply his or her knowledge of human-human-communication.

In natural communication, knowledge about the *context* of the communication is necessary. One part of contextual information is the knowledge about the current *topic*, which can be used for many communicational tasks, e.g. anaphora resolution, managing background knowledge, etc.

Since a household robot needs to be able to adapt to new and changing situations (“open-endedness”), it is not sufficient to use predefined topics. Thus, we decided to build a robot being capable of learning new topics by analysing

dialogues online. To be able to develop and evaluate such a system, we built a corpus containing the relevant information the system could acquire during a communication.

In contrast to the corpus created by (Green et al., 2006) which focuses on interactive aspects such as communicative problems and spatial relations during communication, the BITT corpus is strongly focused on capturing higher dialogue structures (i.e., topics) emerging during human-robot interaction. The corpus is designed to deliver preprocessed data for topic tracking algorithms², facilitating the development and evaluation of such algorithms without an online robot system.

2. Corpus design

In order to get the information a mobile robot system could acquire during a communication, we decided to record the corpus from the robot’s perspective. To be useful for our research, the corpus had to contain natural – not artificially constrained – communication sequences. Additionally, the corpus should contain rich situated topic information.

The setting of the corpus is a so-called Home-Tour scenario, during which a subject introduces a robot to a household-like environment. The advantages of such a scenario for our task are:

1. Home-Tour scenarios resemble basic communication situations for a household robot.
2. The topical structure of the communication is mainly controlled by the subject and not by the robot or both communication partners.
3. The topical structure of the communication is likely to reflect the physical structure of the experimental setting, making it possible to enforce a rich topic structure by the design of the setting.

¹(Cogniron, homepage)

²for example, cf. (Allan, 2002)

2.1. Hardware

For the recording of the corpus we used the mobile robot BIRON³ as a platform (cf. Fig.1).



Fig.1 - BIRON

BIRON is a modified ActivMediaPeopleBot. It is able to detect people, communicate by spoken language, understand simple pointing gestures and detect specified objects. It has the ability to show several facial expressions on its display, thus communicating different states of attention and reporting communication problems in a simple way. Its sensory equipment consists of:

- stereo microphones
- a pan-tilt camera for face tracking
- a laser range scanner
- a stereo camera

We recorded each sensory source except for the laser range scanner, which is only used for person detection tasks, but not for higher dialogue functions. The data gathered from the laser range scanner was used for person tracking during the experiments, though.

Additionally, we recorded the experiments with an external camcorder connected to a headset microphone. This facilitated the manual postprocessing – especially the transcription – of the corpus.

2.2. Experimental design

We invited 29 people to show a specially prepared room to the robot BIRON. The subjects were told that the robot needed the information to introduce the room to seven year old children afterwards. We chose this scenario in order to bias the subjects to elaborate more on the contents of the room. In several cases the people directly instructed BIRON to instruct the “children” about dangerous objects or things to play with, indicating that this approach was successful.

³cf. (Haasch et al., 2004)

2.2.1. Setting

As can be seen on Fig.2, the room contained several topical areas.



Fig.2 - Part of the setting

Examples for topical areas are a kitchenette, a working place, a place to have a cup of tea, etc. In most cases, the topical areas were adopted by the subjects during the description of the room. Only in a few cases, topics spanning the topical areas were developed, for example a topic concerning all the plants in the experimental room.

2.2.2. Robot behaviour

During the experiments, the attention system of BIRON was activated. Thus, it simulated attention by movements of the pan-tilt camera and rotation of the base, tracking the subjects' bodies and faces (cf. Fig.3). The display showed different facial expressions depending on the attention states, e.g., listening or waiting.



Fig.3 - Human-robot interaction example

In order to not restrict natural communication of the subjects by a restricted dialogue system or speech recognition errors, we deactivated the verbal communication capabilities of the robot. This way it behaved only as a listener, but simulated attention by the above mentioned robot reactions. Although the missing verbal feedback of the robot could be estimated as a drawback, we decided not to carry out a Wizard-of-Oz style experiment in order to avoid subject biasing. Also, it is not absolutely clear what tasks future

interaction robot systems will be able to accomplish, so the definition of a simulation of the robot's (i.e., the wizard's) capabilities would have been somewhat arbitrary.

The resulting monologues were an ideal starting point for our research, because they contained free, unconstrained speech, only user-initiated topic shifts and only few disturbances because of robot errors.

3. Postprocessing

Based on hypotheses of cues relevant for topic tracking, we were especially interested in the following information:

1. Pauses (indication of topic shifts)
2. F0 information (indication of topic shifts).
3. Lemmatised spoken language (indication of topics)
4. References to objects or groups of objects by language, gestures or both (indication of topics in situated communication)
5. For the evaluation task we additionally needed information about topics annotated by humans.

Note that except for 5., each of these types of information are or will be available to BIRON during a communication. However, for the corpus we simulated most of the processing steps by manual annotations, to reduce the error rate and get an optimal base for Topic Tracking.

4. Annotation

According to the stated requirements, we decided to include all the data – except for the F0 information – in one single XML-format. For each monologue, phases of spoken language were detected by the automatic voice activity detection⁴ (VAD) we use on BIRON. The monologues were thoroughly transcribed and afterwards enriched by information on multimodal references to objects or object groups, based on gestures, reference by language or both. However, note that no explicit gesture annotation was performed. Only objects that were referenced by deictic gestures were marked in the corpus.

4.1. Time and pause information

The main elements of the corpus are communication segments, i.e., utterances. Each utterance is a continuously spoken part of language which is assumed to bear no topic shifts, i.e. a single – or no – topic. As mentioned above, the utterances were detected using an energy based VAD system. Each utterance bears a start and an end attribute in the corpus files. This way, the start and end attributes of consecutive utterances define speaker pauses. For example:

```
<utterance start="35.11" end="36.39"> (...) and a lot of things (...)</utterance>
<utterance start="36.63" end="41.22"> (...) this desk for example (...)</utterance>
```

Example 1. - Time Information

We found that pause length is a useful indicator of topic shifts in the corpus.

⁴The VAD system is part of the ESMERALDA toolbox, see (Fink, 1999)

4.2. F0 Information

The F0-information was recorded in time-aligned text files generated by PRAAT, cf. (Boersma and Weenik, 2004). We decided to analyse the data gained by the headset microphone instead of the robot microphones because of sound quality reasons.

4.3. Transcription

As mentioned above, the utterances of the subjects were graphemically transcribed. We used a simple annotation scheme to mark aborted utterances or unusually pronounced words. Noise, such as breathing of the subjects, was annotated for speech recognition purposes, but this information was deleted from the topically annotated version of the corpus.

A fixed list of hesitations was defined and used.

For most topic tracking and information retrieval tasks, lemmatisation or stemming is a necessary preprocessing step⁵. In the BITT corpus each utterance is specified as well in a lemmatised as in a not lemmatised form. The given lemmatisation of the BITT-corpus was performed with TreeTagger⁶.

4.4. Reference solution

One of the key aspects of the corpus is the annotation of object references. This information can be used to disambiguate object references from different topics which bear the same verbal description, e.g., “the plant”, but refer to different plants in different contexts.

4.4.1. Object resolution

The annotation differentiates between references to object groups and single objects. Objects were given an ID. References to objects outside the room were not annotated.

Verbal object references were annotated as well as gestural or combined references, although no difference in the annotation scheme was made. In case of verbal references, the referring NP was annotated (cf. Example 2.1). In case of purely gestural object references, the part-of-speech tags were left empty (cf. Example 2.2)

```
there is <object><reference oid="refrigerator_01"/>
<pos> a fridge </pos> </object>(...)
```

Example 2.1 - Verbal object reference resolution

```
look <object><reference oid="plant_02"/> <pos/>
</object>
```

Example 2.2 - Gestural object reference resolution

Purely gestural object references were very rare in the corpus.

4.4.2. Object group resolution

We decided to differentiate between three types of object groups. This distinction was made in order to reduce the object space but to be able to consider references on object groups as well. The three types of object groups are:

⁵cf. (van Rijsbergen, 2005)

⁶see (Schmid, 1994)

1. Groups consisting of items which are almost always⁷ referred to as one single group (e.g., a set of chocolate bars)
2. Abstract groups (e.g., all humans)
3. Groups consisting of individual items which are sometimes referred to as single items and sometimes are mentioned as a part of a group (e.g., a group of plants and the subject refers to the single plants afterwards).

Groups of type (i) and (ii) were treated as single objects, i.e., they were assigned an id:

(...) **if they want to have** <objects><group oid="g_sweets_01" /> <pos>**them**</pos></objects> (...)

Example 3.1 - Object group reference resolution

Groups of type (iii) were annotated by specifying the ids of the objects contained:

(...) **here two** <objects><group oid=""/><member oid="plant_05"/><member oid="plant_01"/><pos>**flowers**</pos></objects> (...)

Example 3.2 - Object group reference resolution

4.5. Topic annotation

Three trained annotators annotated the topic for each utterance. Each annotator had to decide on a non-hierarchical set of topics before beginning the annotation work, but after familiarising him- or herself with the data. Thus, mainly global topics, i.e. topics occurring in more than one monologue, were annotated. We intended to annotate mainly global topics, because for our applications, the detection of local topics is only of limited use: Global topics bear information about recurring events, tasks etc., while local topics do not.

Each annotator assigned one or no topic to each communication segment – communication segments without any topical indication, e.g. “what next”, were intentionally not annotated. This way, the complete corpus was annotated by each annotator, yielding three different topic annotations.

5. Statistics

The BITT corpus consists of 29 transcribed and annotated monologues of 24 female and 5 male subjects, as well as the recorded data from the mentioned sensory sources. The total recording time was 320 minutes, with a mean of about 11 minutes per monologue. The language of the experiments was English, although most of the subjects⁸ were non-native speakers.

2620 references to single items were annotated, as well as 1419 references to groups of objects. The corpus consists of 11209 communication segments. On average 4900 communication segments were annotated with a topic by each annotator.

⁷i.e., in the context of the corpus

⁸There were 2 native and 27 non-native subjects. Each of the subjects estimated his/her English speaking capability as good or better.

6. Outlook and Availability

The BITT corpus is currently used in the development and evaluation of the Topic Tracking software of BIRON. The parts of the corpus which are freely (no fee) available contain:

1. The annotated transcriptions as described above.
2. The time aligned pitch analysis files.

Other parts of the corpus are protected by German data protection law and can only be made available under special circumstances.

If you are interested in using the BITT corpus for research purposes, please contact us by one of the email addresses given in the header.

7. Acknowledgments

This research is supported by the Graduate Program “Task Oriented Communication”, funded by the German Research Foundation, and by the COGNIRON-project (FP6-IST-002020), supported by the European Union.

8. References

- James Allan (ed.) 2002. *Topic Detection and Tracking*. Kluwer Academic Publishers, Norwell, Massachusetts.
- Paul Boersma and David Weenink 2004. Praat: doing phonetics by computer (Version 4.2.17) [Computer program]. URL: <http://www.praat.org/>
- COGNIRON - The Cognitive Robot Companion – Project homepage. URL: <http://www.cogniron.org/>
- A. Green, H. Hüttenrauch, E. A. Topp and K.S. Eklundh 2006. *Developing a Contextualized Multimodal Corpus for Human-Robot Interaction*. International Conference on Language Resources and Evaluation (LREC), Genua.
- A. Haasch, S. Hohanner, S. Hüwel, M. Kleinhagenbrock, S. Lang, I. Tóptsis, G. A. Fink, J. Fritsch, B. Wrede and G. Sagerer. 2004. *BIRON – The Bielefeld Robot Companion*. E. Prassler, G. Lawitzky, P. Fiorini and M. Hägele (ed.): Proc. Int. Workshop on Advances in Service Robotics, pp. 27–32.
1999. G. A. Fink: *Developing HMM-based recognizers with ESMERALDA*. V. Matousek, P. Mautner, J. Ocelkov, and P. Sojka (ed.): Lecture Notes in Artificial Intelligence, volume 1692, pp. 229–234, Berlin Heidelberg. Springer.
2005. C.J. van Rijsbergen: *Information Retrieval*. URL: <http://www.dcs.gla.ac.uk/Keith/Preface.html>, 2nd edition.
1994. H. Schmid: *Probabilistic Part-of-Speech Tagging Using Decision Trees*. International Conference on New Methods in Language Processing.

Towards a Multimodal Topic Tracking System for a Mobile Robot

Jan F. Maas, Britta Wrede, Gerhard Sagerer

Applied Computer Science Group
Faculty of Technology, Bielefeld University, Germany
{jmaas,bwrede,sagerer}@techfak.uni-bielefeld.de

Abstract

Topics in situated and task oriented communication depend heavily on the given, often changing environment, making the detection of predetermined topics in many cases useless. Detection of non-predefined topics can enhance Human-Robot-Interaction (HRI) in a variety of ways, though.

In this paper we propose a way to dynamically determine topics during Human-Robot-Communication using well established techniques such as Latent Semantic Analysis (LSA). The procedure is based on multimodal cues, supporting the view that topics are not simply a property of spoken or written language, but of multimodal situated communication. An online version of the topic detection system has been developed and is currently being tested on our mobile robot BIRON. To demonstrate the feasibility of our approach, we present the results of an evaluation of our system on the BITT corpus.

Index Terms: topic tracking, multimodal dialogue, human robot interaction.

1. Introduction

During the last years, topic tracking has been used to support many different tasks in natural language processing. For example, topic tracking techniques have been developed in the context of information retrieval [1] to make relevant information from written (e.g., newspaper) corpora accessible. Also, topic detection and tracking techniques have been developed for broadcast news [2] or spoken language databases [3], thus allowing topic sensitive search on broadcast news databases or building intelligent radios, which scan for selected information.

In our work we want to make the benefits of TDT in situated, task-oriented Human-robot-interaction (HRI) accessible. The goal is the design of a system which allows a robot to track topics during communication. The knowledge about the relevant topic can support several tasks of the robot, such as speech recognition, speech understanding or social interaction. For example, speech recognition can be enhanced by adjusting the language model according to the current topic, while speech understanding can be enhanced by resolving underspecified references: The sentences “My shoes are dirty. Fetch me a brush” do not specify the kind of brush. Topical knowledge can help determining that a shoe brush is asked for.

Generally speaking the TDT information can be used for adaptation of the robot to a given environment, allowing “soft” adaptation rather than learning “hard” facts, i.e., propositions.

1.1. Scenario

The research is embedded in the framework of the COGNIRON-project [4]. The main goal of the project is to study perceptual,

representational, reasoning and learning capabilities of embodied robots in human centred environments, e.g., private homes.

Within this framework, several scenarios were defined to obtain simplified access to the complex situations that can occur. Our TDT research is focused on the so called “home tour” scenario where a robot is introduced to an unknown private home environment. The robot should be able to communicate by means of spoken language but also be able to make use of multimodal information as well. It should be able to follow users, ask questions and learn about the environment.

1.1.1. Comparison to other scenarios

In comparison to other application areas of Topic Spotting or TDT, several differences are obvious. A difference which makes topic detection in situated communication difficult concerns the size of the entities which are to be classified: A newspaper article usually has more than hundred words, containing many features allowing a classification. In HRI it is possible to have multiple topic changes during a single dialogue turn.

Another challenge is the necessity to detect *dynamic* (not pre-defined) topics, emerging through the structure of the environment. E.g., it is not useful to predefine topics such as “living room” for there may be none, whereas a billiard room may exist which is very unlikely in most homes and thus to appear in a list of pre-defined topics. These circumstances make solely word-based dynamic TDT in situated HRI nearly impossible. To cope with the difficulties we propose the use of multimodal information such as gestures, object references, identification of communication partners, etc., which will be available to our robot system.

1.2. Related work

While the detection of topics in situated, multimodal HRI is a new area of research, our method makes use of former successful approaches of topic detection. The structure of the system roughly follows the tasks defined in the Topic Detection and Tracking project [2]. A related approach of dynamic topic detection on spoken language was developed by [3]. The capability of Semiotic Spaces – which are the basic technique of our topic tracking approach – to cope with additional, dialogue related information was shown in [5]. Our approach of segmentation of spoken language into topics made use of the results in [6].

While most of these approaches have been proved to work on large non-situated corpora, it is our goal to develop a strategy for topic tracking in situated communication with few training data supported by multimodal information.

2. Topic Tracking Approach

Before we explain the method underlying our approach, we want to define the notion “topic”. Intuitively, a topic is “what a text or discourse segment is about”. It is possible to have several names for a topic, e.g., “washing clothes” or “how to do the laundry” etc. ([7], p 72), therefore we do not indicate topics by speaking names but by sets of signs which are thematically related (e.g., “washing machine”, “washing powder”, “water”, “tap” etc.).

In our approach we define a **topic** τ as an abstract property of communication segments. A **communication segment** is a collection of signs perceived by a robot during a certain period of time. For example: A human shows the task of dish-washing to a robot. The words uttered by the human, but also the objects involved in the interaction and possibly more signs build a communication segment.

A communication segment is assumed to bear exactly one topic, although several consecutive communication segments may have the same topic. Merging such communication segments would result in another, larger communication segment.

A perceived sign, which is an element of a certain communication segment, may indicate the topic of the communication segment. For example, the word “fridge” uttered during a communication segment may indicate a topic centred around a kitchen. Also, the gestural reference to the fridge can do the same.

2.1. General Idea

It is our goal to dynamically build and detect topics emerging during a human-robot-interaction. To reach this goal, we gather information about signs commonly co-occurring during communication segments, which are derived from a continuous communication during a segmentation process. Signs which commonly co-occur are assumed to belong to the same topic. This way, the signs of former communication segments can - if encountered again - be used to recognise the current topic. The information gathered about this topic so far can then be used for different tasks.

In our approach, a topic is represented by the set of all topic indicating signs (e.g., function words – like “and” – which indicate no topic are not part of this set) and the indication strength of the sign for the topic (see below). This way, a simple weighted unigram classification can be carried out to detect the topic of newly encountered signs, i.e. communication segments.

2.2. Why using Semiotic Spaces?

The idea of representing semantic relationships by utilising co-occurrence information is the central idea of Semiotic Spaces [8]. In information retrieval research Semiotic Spaces have been successfully used to build up groups of texts or words which are similar or belong to the same topic [9]. Semiotic Spaces are not restricted to words or a single source of information, but can be applied to data from mixed sources of information as well [5]. Additionally, Semiotic Spaces can be applied to very small training sets, thus allowing for topic tracking in very short communications. Further, they are robust against dialogue parsing errors, since they rely on symbol-based cues. For these reasons, we decided to use Semiotic Spaces as the core technique for our research.

2.3. Types of Semiotic Spaces

Semiotic Spaces are in most cases high-dimensional vector spaces into which texts or words are projected. In these spaces, the dis-

tances of the word or text vectors are a measure of similarity, similar entities being close to each other. The more often signs co-occur, the smaller the distance between the signs will become.

In text technology several kinds of Semiotic Spaces exist [8]. The Semiotic Spaces differ in their power to represent semantic associations and their computational speed, but additionally in the need for specifying the number of topics beforehand. For this reason, we did not use Support Vector Machines (SVM) or Probabilistic Latent Semantic Analysis (PLSA), because the highly dynamic household environments do not allow to specify the number of topics in advance.

For our experiments we employed three Semiotic Spaces: The simple vector space model [10], Fuzzy Semantics [11] and LSA [9]. In the simple vector space model, each sign is represented by a vector $v = (d_1, \dots, d_n)$, with d_n being the entropy-weighted frequency of the sign in the c 'th communication segment of the training set. Fuzzy Semantics spaces are normalised simple vector spaces, while LSA employs dimensionality reduction on the data to reduce the impact of “noise”.

2.4. Procedure

The topic tracking procedure consists of five steps in order to compute the indication strength for each topic indicating sign: Segmentation, preprocessing, semantic association, clustering of the Semiotic Space and estimation of support strength. These steps are computed on training data. Segmentation and preprocessing are also preparatory steps for the Topic Tracking on new communication segments. Depending on the hardware and algorithms used, each new communication segment can be employed for training after its classification, enabling topic learning during the communication.

1. Segmentation: The goal of the segmentation is to split a communication process into communication segments, which ideally have exactly one topic. During the segmentation phase, signs from different modalities are merged in a “bag of signs” which constitute the communication segment. The segmentation process is the main information source for the unsupervised learning task, which takes place during the calculation of the Semiotic Space and clustering. While segmentation is an important issue for our topic tracking approach, it heavily depends on the kind of communication. For our system, we use several multimodal cues such as movement of the user, eye (camera) contact and dialogue structure.

Consider that two different segmentation processes take place for training and tracking: While the segmentation process for training results in long communication segments to gather as much co-occurrence information as possible, the segmentation process for the tracking delivers short pieces of continuous communication, e.g., utterances. This is done to allow real-time tracking. To avoid confusion, we will call the segments which result from the first method “communication segments”, and the segments for tracking “chunks”. Further, the notion “segmentation” will only be applied to the process of creating communication segments. Chunks are gained by an automatic voice activity detection both for the online system and the offline experiments.

2. Preprocessing: Preprocessing is concerned with deleting words that are unlikely to indicate topics, e.g., function words, not conventionalised gestures, etc. Also, signs may be transformed into more abstract signs when they have the same relevance for a topic (e.g., by lemmatisation). In our system, a stoplist of function words was manually created. Lemmatisation is also done by a manually edited lookup table, although more elaborated algo-

gorithms exist and were utilised for the offline experiments (see below).

3. Semiotic Space: This step is needed to compute the distances of signs to each other, building the necessary foundation for a topic clustering. Distances can be computed with several standard distance measures such as cosine, Euclidean distance, etc. In our system, we usually apply correlation coefficients.

4. Clustering: The Semiotic Space only represents thematic distances between signs. To gather information about interrelated signs, i.e., topics, a clustering algorithm is applied. For our purposes, we use average-linkage agglomerative clustering [12], since it is fast and allows for dynamic segmentation by specifying a clustering limit. During the experiments, the algorithm was constrained to detect 10 clusters or less – slightly more than in the manual annotation – to support comparability.

5. Classification: Preceding classification, the average distance of each sign to each sign within a topic cluster is computed. To determine the topic of a chunk, we simply sum the distances of the signs in the chunk for each topic, thus implementing a weighted unigram classification. The topic with the highest value is the detected topic. Should no topic be detected because of no topic indicating signs, the last detected topic is registered.

Consider that in the online system, steps 3 and 4 are continually repeated during communication, thus updating the system. For offline evaluation, topic clusters based on a training set were created and not modified during tracking (see below).

3. Corpus

Evaluation of our topic tracking system was performed on the BITT-corpus [13], which was developed for evaluation of topic tracking on mobile robots. The corpus consists of 29 human-robot interactions in a scenario similar to the home tour. The subjects introduced a room they had familiarised themselves with to the robot BIRON [14]. The room was specially prepared to contain a lot of items which topically belong together, for example a place to play games, a kitchenette, a working place, etc. Thus we hoped, different topics related to the different areas would occur naturally during the communication.

BIRON is a modified Pioneer PeopleBot from ActivMedia. It was able to turn the base during the experiment, to watch the instructor and to track the face of the user with a pan-tilt camera. Thus, the robot simulated attention. No more robot reactions took place in order to minimise communication problems, allowing the people to communicate most naturally and without special instructions. Because of the small number of constraints the resulting data was highly heterogeneous: People used different communication strategies, different detail levels, etc. The duration of the experiments varied from 3 minutes up to over 30 minutes, with about 9.5 minutes average. The interactions were recorded and manually transcribed.

Annotation After the transcription, object references were annotated, thus simulating the capability of our robot to detect objects which were verbally and/or gesturally referred to. Each object was given an ID. In some cases, object groups were marked, e.g., a set of chocolate bars which were usually referred to as a set were given a single ID. Consider that almost no gestures referring to objects occurred without accompanying speech, although the gestures were in many cases necessary to resolve the reference.

The signs (words and object references) were grouped in short intervals of continuous speech – the chunks which are classified

during tracking – by BIRON’s voice activity detection. Time information (start and end) was added to each chunk.

To facilitate topic tracking evaluation, three annotators noted a topic for each chunk. The annotators decided on the topics for themselves before annotation and after making themselves familiar with the corpus. This way, dynamic topic tracking on the corpus was performed manually.

4. Evaluation

For evaluation, each of the 29 monologues of the BITT corpus was tracked by our system. The training set for each monologue consisted of the 28 other monologues. Lemmatisation of the words of the corpus was performed with the TreeTagger [15].

To create communication segments for building the training sets, subsequent chunks bearing the same manually annotated topic from the topic annotation of the BITT corpus were merged.

4.1. Evaluation metric

We compared the results with the three manual topic annotations of the BITT corpus. However, a direct comparison is not possible because of two reasons: First, the number of automatically and manually detected topic may change from communication to communication. Second, a decision process is needed for determining which dynamically detected topic represents which manually detected topic.

To cope with these problems, we developed the following evaluation metric:

1. The chunks which were automatically classified as belonging to a single topic were determined.
2. For these chunks, the quantity of chunks belonging to each manually determined topic was determined.
3. The chunks bearing the manually detected topic with the highest count were considered as correct, the chunks bearing other manually detected topics were considered as false.

This process was repeated for each automatically detected topic, the results were summed up. With nearly 4900 manually annotated chunks for each of the three annotators, each evaluation is based on about 14.500 single values.

The topic tracking system was biased to detect about as many topics as were determined by the human annotators by specifying a maximum number of topics to be detected.

4.2. Results

We now want to discuss the results from the offline evaluation process. Since we believe that for topic tracking on situated HRI a multimodal approach is needed, we varied the available sign types for training. The system was tested with words (Table 1), with verbal object references replaced by object IDs (Table 2) and with object IDs obtained from verbal and/or gestural references only (Table 3). Also, the amount of signs used for each evaluation was varied by setting C_{min} , the minimum number of communication segments in the training set in which a sign has occur to be considered. While testing the mixed approach, words were given a 50% penalty on their topic indication strength compared to object references. This was undertaken to enhance the results, since words were assumed to be less reliable in topic indicating than object references.

Table 1 shows the results for only word-based tracking. Although the results are not satisfying, three tendencies are observable: First,

Table 1: *Word based approach*

CMin	2	3	4	5	7	10	15	20
LSA	64.7	64.2	64.0	64.2	64.3	62.3	61.1	57.2
FSem	58.3	60.4	59.8	61.6	63.0	60.5	59.1	53.8
Vec.	50.1	56.4	57.6	61.7	58.9	60.6	59.4	53.5

LSA performs better than the other approaches, with the simple vector space model as the worst. Second, the approach based on the simple vectorspace is somewhat susceptible to disturbances by rarely used words. With such words deleted, the performance of the model increases. Finally, the performance of each model drops when too many topic indicating cues are restricted, i.e. for high values of C_{min} .

Table 2: *Mixed approach, penalty of 50% for words*

CMin	2	3	4	5	7	10	15	20
LSA	80.6	81.5	82.6	81.2	81.0	80.3	73.6	63.8
FSem	69.4	72.8	69.9	66.5	71.1	69.9	73.8	64.6
Vec.	62.4	72.0	74.3	80.4	79.8	79.3	72.4	62.5

The same tendencies can be found in Table 2, where the simple vector space model sometimes performs even better than the Fuzzy Semantics approach. The fluctuation in the results of the Fuzzy Semantics Space are probably an indicator for the limited capability of the model to incorporate cues from modalities which share different distributions in the corpus. For topic tracking in HRI, it is desirable to cope with small values for C_{min} , since more information is gained and smaller training sets are necessary.

Table 3: *References only*

CMin	2	3	4	5	7	10	15	20
LSA	89.4	90.2	89.1	88.8	87.7	87.0	81.9	71.2
FSem	86.2	89.2	89.0	89.4	88.7	88.0	81.4	70.9
Vec.	90.0	90.0	88.9	90.0	87.4	86.2	82.1	71.3

Table 3 finally shows that topic tracking based only on multimodally resolved object references performs best on the corpus, indicating a strong connection of topics and objects in the environment. Interestingly, the three models show no great differences in performance, probably because of the fact that objects belong to topics nearly on a 1:1 basis, thus reducing “noise” tremendously.

5. Discussion and future work

The results show that using multimodal information is necessary for topic tracking on situated HRI. Obviously, object references are strong topic markers. The mixed model shows acceptable performance while delivering additional information compared to the reference based approach, namely word/topic relationships. Since this information is desirable, future work will comprise the development of mixed approaches, combining the good tracking results of solely object-based approaches with the additional information of mixed models. Solely word-based-tracking performs poorly, thus supporting our assumption.

Based on this offline system, an online topic tracking system for the robot BIRON has been developed. At the moment, experiments concerning topic tracking during communication sequences with BIRON are carried out. During this evaluation, the automatic segmentation will be tested. Future work will especially focus on

the further development, integration and enhancement of the topic tracking software for the mobile robot.

6. Acknowledgements

The research has been supported by the European Union within the “Cognitive Robot Companion” (COGNIRON) project (FP6-IST-002020) and by the German Research Foundation within the Graduate Program “Task Oriented Communication”.

7. References

- [1] C.J. van Rijsbergen: “Information Retrieval”. 2nd edition. URL: <http://www.dcs.gla.ac.uk/Keith/Preface.html>. 2005.
- [2] J. Allan (ed.): “Topic Detection and Tracking”. Kluwer Academic Publishers, Norwell, Massachusetts, 2002.
- [3] Mikko Kurimo: “Thematic Indexing of Spoken Documents by Using Self-Organizing Maps”. in: Speech Communication. Vol. 38. pp. 29-44.
- [4] COGNIRON - The Cognitive Robot Companion. Project homepage. URL: <http://www.cogniron.org/>.
- [5] R. Serafin and B. Di Eugenio: “FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification”. ACL, New York, 2004. pp. 692-699.
- [6] E. Shriberg, A. Stolcke, D. Hakkani-Tür and G. Tür: “Prosody-Based Automatic Segmentation of Speech into Sentences and Topics”. Speech Communication, Vol. 1-2, 2000, pp. 127-154.
- [7] G. Brown, G. Yule: “Discourse Analysis”. Cambridge University Press, Cambridge, 1983.
- [8] E. Leopold: “On Semiotic Spaces” in: Christian Wolff (ed.): LDV-Forum 18 (3). 2005.
- [9] T.K. Landauer and S. T. Dumais: “A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Aquisition, Induction and Representation of Knowledge”, Psychological review, 1997, Vol. 104, pp. 211-240.
- [10] G. Salton: “Automatic Information Organization and Retrieval”. McGraw-Hill Book Company, New York, 1968.
- [11] B. B. Rieger: “Connotative Dependency Structures in Semiotic Space” in: Rieger, B. (ed.): Empirical Semiotics II. Quantative Linguistics. Vol. 13, Bochum 1981, pp. 622-711.
- [12] P. Cimiano, A. Hotho and S. Staab: “Comparing Conceptual, Partitional and Agglomerative Clustering for Learning Taxonomies from Text” in: Proceedings of the European Conference on Artificial Intelligence (ECAI’04), 2004, pp 435-439.
- [13] Jan F. Maas, B. Wrede: “BITT: A Corpus for Topic Tracking Evaluation on Multimodal Human-Robot-Interaction”. Proceedings of the international conference on Language and Evaluation (LREC), Genoa, Italy. 2006. (to appear).
- [14] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinhagenbrock, S. Lang, I. Tóptsis, G. A. Fink, J. Fritsch, B. Wrede and G. Sagerer: “BIRON – The Bielefeld Robot Companion” in: E. Prassler, G. Lawitzky, P. Fiorini and M. Hägele (ed.): Proc. Int. Workshop on Advances in Service Robotics, 2004, pp. 27-32.
- [15] H. Schmid: “Probabilistic Part-of-Speech Tagging Using Decision Trees” Proceedings of the International Conference on New Methods in Language Processing. 1994.

Interaction between dialog, supervision and planning

Joint RA6-RA1 working document

January 2007

1 Introduction

This document reports on several work meetings and discussions oriented toward the incremental design of the interplay between the robot supervisor called “SHARY” (Supervisor for Human Aware Robot Ynteraction”) , the high-level planner called “HATP” (Human-Aware Task Planner) and the Dialog system.

The study has essentially focused on the verbal and non-verbal communication acts associated to the establishment of a joint task between a robot and its human partner and the different decisions that are involved in the task achievement.

This is a joint between RA 1 and RA 6.

2 The framework

The chosen framework is the joint activity paradigm. In the case of Cogniron, we shift from agent-agent to human-robot interaction and more particularly here on interactive/collaborative task achievement.

The framework and the associated requirements that have been expressed above involve various issues in Cogniron. We concentrate here on the interaction between Supervision (and the planner as a resource) and Dialog.

We would like to establish mechanisms for flexible coordination and turn-taking in a context where the robot will try to acquire human acquiescence of its behaviour.

This will be done on the basis of concrete problems borrowed from KE2 and KE1.

The starting point is that we envision KE2 and KE1 as relevant context for joint activity.

This might be present at several levels of the task hierarchy. For instance, in KE2, when the human asks for a drink, the overall goal is treated as a joint activity between the human and the robot. During execution, several (sub-tasks) will be performed. Some sub-tasks will be treated as individual tasks (e.g. when the robot picks an objects or navigates to a given place) and some sub-tasks will be treated as joint tasks (e.g. when the robot hands the can to the person).

For KE1, it is interesting to study the overall home tour as a joint activity where the human and the robot “collaborate”. the overall goal is achieved when the robot “knows enough” about the home. During this overall process, sub-tasks can be also analysed from this point of view, for instance when the human points and names a object. Indeed, a complete transaction should be performed where the robot and the person are involved: motion of the robot in order to place itself in a proper position, perception of the object and possible iterations of the process supported by dialog, etc..

2.1 Possible communicative acts

At any time, both the user and the robot can propose the following task-based communication acts:

ASK_TASK proposing a task.

PROPOSE_PLAN proposing a plan (recipe) for a certain task

PROPOSE_MODIFY_PLAN proposing a modification of the current plan for a certain task

GIVE_UP gives up a task (e.g., because the task becomes impossible). For the robot this is a way to announce that it is unable to achieve the task.

CANCEL cancellation of a task (voluntary give-up)

TASK_DONE announces that the task has been done

REALIZE_TASK announces that the task performance will start.

The dialog negotiates with the user till she provides a clear reply which the dialog can associate with a set of pre-defined supervision requests. This means that the supervision will always get a clear reply of the user.

This process is performed along the task hierarchy that is produced by the planner or encoded as procedural knowledge in the supervisor (Open-PRS provides a context-dependent goal decomposition).

Supervision and dialog interact about the SuperTask and joint subtasks, but not on sub-tasks that involve only the robot.

2.2 Communication context

One or two steps: Depending on the situation an answer may follow directly a request or not. In such a case, there is an acknowledgement and the answer come later.

Explicit or implicit communication Note that, depending on the context, the communication acts may be implicit. For example in some situations, the robot will proceed without asking or proposing a plan/recipe because it considers that it is in a situation where its intentionality is “visible” and where the human is supposed to know what the robot is doing. Other examples are the monitoring of situations that can be interpreted as a signal of CANCEL for instance (e.g. if the person retracts his hand during while the robot is handing him an object).

Remark on “public” and “private” plans: planning and supervision system consider plan at two levels, i.e. a plan got a “public” part that has to be shared with the participant of the task and every modification should be discussed; and a “private” part for which there is no need to inform the partner (ex: robot changes its motion algorithm for a better adapted one). Only the public part of the plan will be communicated to the dialog system.

3 Scenario and variations

The scenario is based on the KE2 scenario.

3.1 Basic scenario

Jido is in its waiting zone. Luis comes and asks Jido: “robot, bring me a coke”. Jido accepts. It goes to the cafeteria, finds the cola can, and brings back the cola to Luis’ office. Jido and Luis exchange the cola can.

3.2 Variation 1

Example 1 Jido meets Luis at the cafeteria. Before it takes the can, the robot proposes Luis to take the can himself. Two possibilities:

- Luis agrees.,
- Luis refuses.

Example 2 Jido meets Luis while heading to Luis Office after it has taken the can. Jido proposes to Luis to give immediately. Two possibilities:

- Luis agrees and takes the cola can,
- Luis refuses.

Example 3 Jido goes to the cafeteria, no cola anymore, but another position is available for the can (and Luis do not said specifically from where the cola should come).

Analysis

A new plan is being constructed during the execution. The information comes from the planning system. The robot proposes it to the user.

Example 1 and 2 consider changes of "public" part of the plan and example 3 considers the "private" part (given that Luis did not say anything concerning the cola location, that would be different in that case).

The dialog system will receive a message asking for a plan modification and will generate appropriate message to the user.

Question : even if a better plan has been computed, do we have to consider to disturb the human every time (that will not happen so many times), or just when it is in a zone "around" the robot, or when the robot "knows" that the human can be disturbed ? This is clearly context and user dependent.

3.3 Variation 2

"Bring me a coffee from the cafeteria."

"Bring me a coffee in my office"

Analysis

In that case, we could consider that user adds constraints to the task.

It seems reasonable that we can have some objects with the same type or many locations possible for a given object.

If the task comes with constraints, we should add constraints to the task definition, so we should define a language of constraints. Constraints will be defined in an independent manner:

TASK + CONSTRAINTS

Constraints will be on instantiation of variables (e.g. : object/person name, place)

The set of tasks and possible constraints associated to them must be made available.

Question is it possible for the dialog system to analyse this type of sentence as: Task : Bring(Coffee, Luis, constraints (Coffee.Location = Cafeteria))

3.4 Variation 3

"Bring me a coffee. Put it on my desk."

"Bring me a coke to my office. Oh, but put it on my desk (not on my chair)."

"Bring me a coffee. Give it to me."

Jido arrives at Luis office, Luis said "put it there please".

Analysis

Here the user gives/imposes a part of the plan.

How does dialog know that these are not two tasks, but one task and a plan proposal / modification? This, clearly cannot be disambiguated if its not linked with the overall joint activity.

Another problem for the dialog is to refer to "it" and my in *"Put it on my desk"* as the coffee and the user desk.

Question for user studies (e.g. by KTH?) : Will users utter more likely:

1. two different tasks in a sequence (e.g. "bring me a coke and then water the flower")
2. or a task and a plan proposal (e.g. "bring me a coke and put it on the table")

Even if it is not possible for the dialog to disambiguate, if the information could be communicated to the supervision level as two consecutive **ASK_TASK**. Another possibility is to have access to the possible decomposition of "bring" in order to detect that "put" is to be interpreted as a sub-task.

3.5 Variation 4

"if I'm not in my office, put it on my desk"

Analysis

Conditional modification of the plan.

Questions

- dialog: is it possible for the dialog to understand the conditional ?
- planning: for now, seems to difficult to take this type of information into account.

This is clearly an interesting problem since the user give a high-level script.

3.6 Variation 5

No plan is found.

Analysis

Robot can say that no plan has been found but cannot say why. Indeed, standard planning systems today are not able, in the general case, to provide explanations of their failure to find a plan.

3.7 Variation 6

"Bring me a coffee. I will be in my office"

Analysis

In that case, user gives an information that will be true later. Seems not so simple from the planning part to model this type of constraint and to take it into account. However, HATP (under construction) will have the possibility to build plans that synchronize with expected events (temporal planning).

Question concerning the dialog system, is it something understandable ? Seems two separate sentences needed.

3.8 Variation 7

"bring me a coke ah... and bring me a mars"

This has to be interpreted as a conjunction of goals, or as two tasks that do not need necessarily to be solved in sequence....

3.9 Variation 8

Theoretically, when a planner is involved, this could help to process indirect speech acts such as:

U: Oh, it's cold here! No wonder, the window is open.

Thus, the dialog component could enter two facts to the common database: (1) user thinks it's cold (2) window is open. A planner might be able to deduce from this that it would be more comfortable for the user if the window was closed. The planner could thus compute a plan for closing the window and send this plan-suggestion to the dialog component. The dialog would then react with:

R: Should I close the windows?

But this requires a common database which doesn't seem to be that simple and thus is rather future work.

3.10 Variation 9

can you do that with that parameters ? what can you do ?

This variation illustrates the difference between questions about the general abilities of the robot and questions related to its capacity to answer if it is able to achieve a task in the current context.

4 Interaction between CF- DLG and CD-RET

We have defined a first sketch of the interaction protocol and data exchanges between Dialog and supervision for human-robot dialog interactions linked to "joint tasks".

The planner intervenes as a "resource" that is invoked by the supervisor, when necessary, to assess the feasibility of a task in a given context.

The examples are borrowed from the use cases: basic scenario and variation 1. Three steps have been analyzed: START (figure 1), TAKE(figure 2) and GIVE (figure 3).

The different steps are decomposed along the axis Dialog/Supervision/Planning.

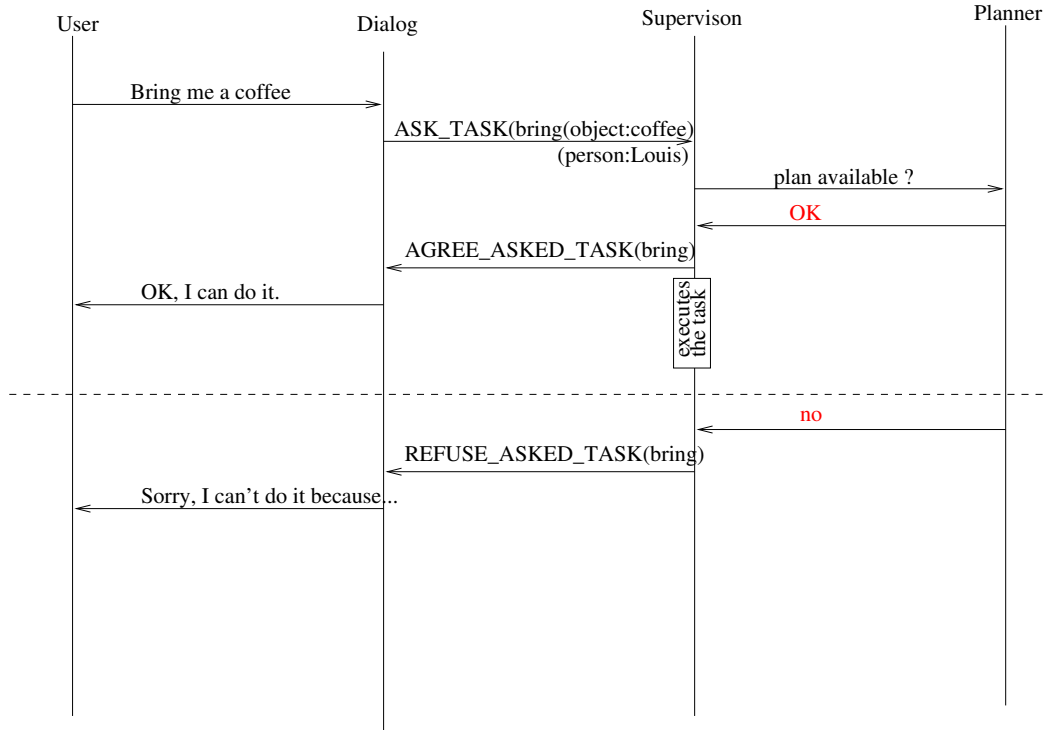


Figure 1: START: the human asks for a task, the robot accepts or refuses depending on it ability to find or synthesize a plan

5 Interface between Dialog and Supervision

We provide here below a preliminary definition of the data structures that will be exchanged between the Dialog and the Supervisor.

5.1 From Supervision to DLG

```

<MSG>
<sender>Supervision</sender>
<receiver>DLG</receiver>
<commAct state = "launch/stop">
<act>the communicative acts above</act>
<property>agree or refuse or act</property>
<info>
<task id =
"the ID of a task (specified by the user) or
a subtask (specified by the planner)">

<person>...</person>
<obj>...</obj>
...
</task>
<plan feature = "new or current">the name of the plan</plan>
...
</info>
</commAct>
  
```

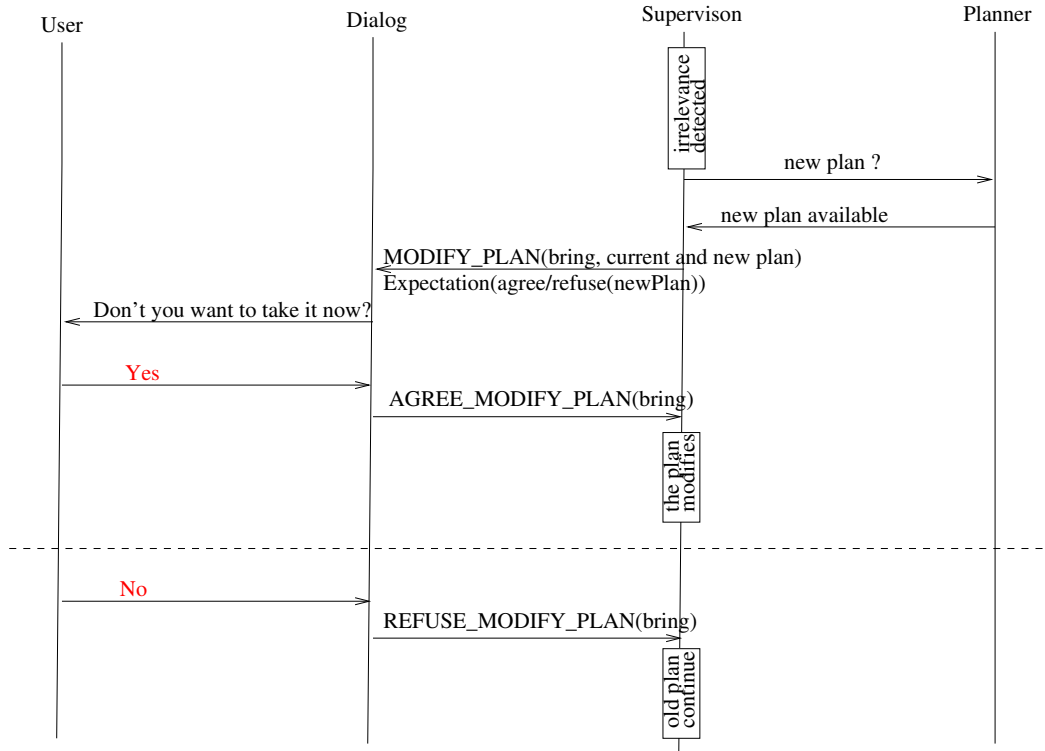


Figure 2: TAKE: the robot encounters the user before reaching the place where it was supposed to give to can to him: it proposes to hand it in the current place. This is a situation where the robot decides a communicative act and proposes a new “recipe”

```

<expectation>
<commAct> as defined above</commAct>
...
</expectation>
</MSG>

```

Explanation :

- `<commAct state = "launch/stop">` :
 - launch: the supervision is launching the communicative act,
 - stop: the supervision just stops this act.
- `<property>agree or refuse or act</property>` : an attribute “act” is the content that is refused or agreed upon

5.2 From DLG to Supervision

```

<MSG>
<sender>DLG</sender>
<receiver>Supervision</receiver>
<commAct state = "final/negotiating">
<act>the communicative acts above</act>
<property>agree or refuse or act</property>
<info>

```

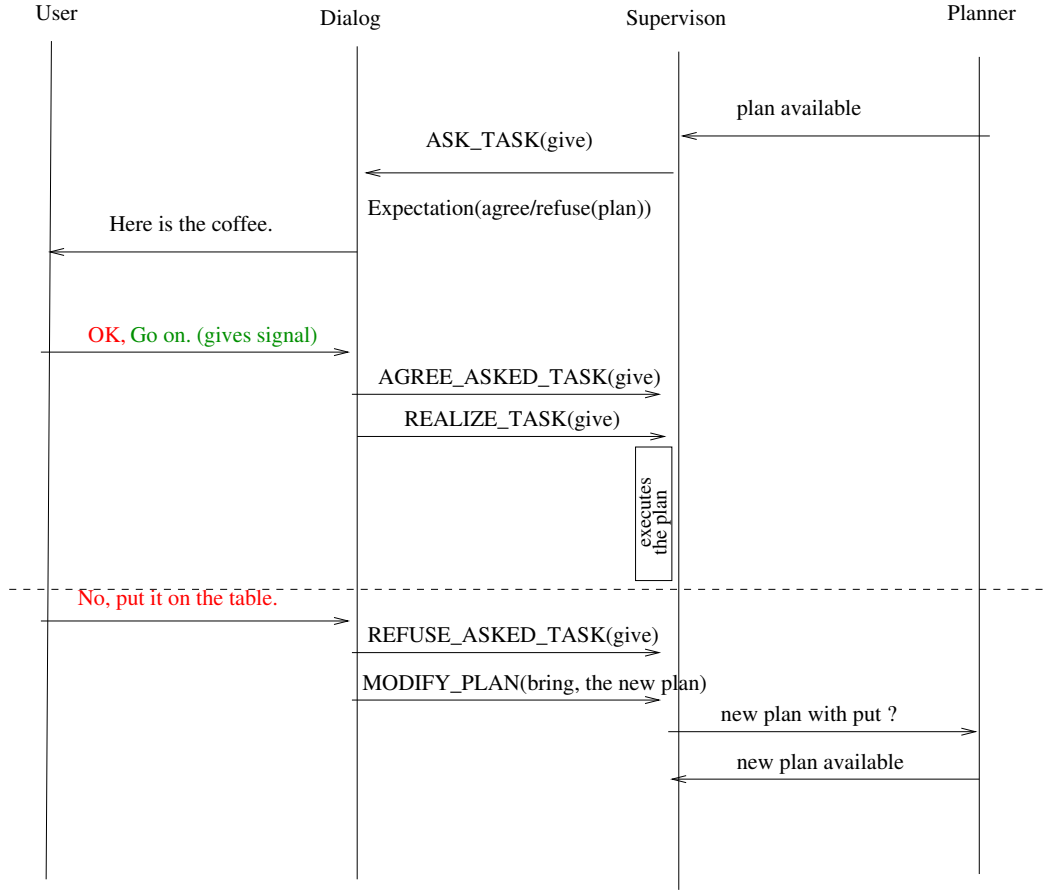


Figure 3: GIVE: the robot hands a coffee to the user. In the first case, the user agrees with the current task achievement and the robot performs its plan. In the second case, the human rejects the plans and proposes an alternative solution to the same joint task: put the coffee on the table.

```

<task>
<person>...</person>
<obj>...</obj>
<position>the concrete position or gestureExpected6</position>
...
</task>
<plan feature = "new or current">the name of the plan</plan>
...
</info>
</commAct>
</MSG>
  
```

Explanation :

- `<commAct state = "final/negotiating">` :
 - final: the final result of the dialog's processing as to what the user said,
 - negotiating: the dialog is still negotiating with the user, in this case, the following commAct may not contain all the information

- `<act>the communicative acts above</act>` : if the dialog knows the commAct, then the value of "act" will be set to
- `<property>agree or refuse or act</property>` : an attribute act is the content that is refused or agreed
- `<position>the concrete position or gestureExpected</position>` : "gestureExpected" is sent to Supervision, if the speech understanding indicates that a gesture is expected, e.g., in case of "this" or "there" and if no gesture result can be found in the gesture-database. After receiving this message, the supervision will activate the gesture recognition whose result will be put in the database. The dialog will access this database to retrieve the result. If there is still no result, then the dialog will continue to negotiate with the user.

User Models Based on Initial Interaction Patterns and Deictic/Ergotic Gestures

Anders Green

Royal institute of technology
100 44 Stockholm
Sweden
green@csc.kth.se

1 Introduction

In the Cogniron corpus (Green et al., 2006) we have collected and annotated several types of data that gives us an opportunity to classify users according to their behavior. In the so called Home Tour scenario described in (Green et al., 2004) the users were asked to *show objects and locations* to a service robot using a combination of speech and gesture. By using the corpus we can analyse interaction between users and robots in terms of communicative behavior and other actions that can be used to group users.

In the robotics scenario we are investigating the adaptation that concerns the actual interaction model, i.e., how we can adapt the interactive behavior of the dialogue system depending on the type of user.

Fischer (2006) used internal criteria to group users into *players* and *non-players*. According to the model proposed by Fischer (Fischer., 2006) users of a task-oriented natural language dialogue system can be grouped into two major groups depending on their responses to the first contribution of the system, e.g., “how do you do?”. One group, referred to as *players* are more willing to play along within the scenario, they are more cooperative and the communication displays properties that show similarities with human-human natural language interaction. We might use this line of reasoning and apply it on our corpus data, trying to find some key phrase, combination of gesture or other behavior that could be used to distinguish between types of users.

- One assumption is that the way objects and locations are introduced to the robot can be used

as a characteristic for grouping users.

- Furthermore the first encounter of communicative problem can provide a cue to whether the user is actively playing along or is primarily using the robot as a tool.

In the following we will examine two types of user activity that we might use to distinguish between types of users. First, in section 1.1, we will look at interaction from an *information point of view* with respect to the specific action sequences the user follows and the results in terms of robot behaviour. Secondly, in section 1.2, we will attempt to establish user types that are based on communicative behaviour as such.

1.1 What do unsuspecting users do?

A user model that is based on the information available to users can be maintained by maintaining an action history, e.g., by storing communicative and other actions. Using this history we can apply heuristics to arrive at an information state for a particular user that can be used to group users.

Again we are interested in establishing differences between users early in the interaction, hence we are looking at *initial interaction patterns*. If we look at the sequence of interaction between the interaction with the robot to the first problem encountered we find the following:

- Every user greets the robot in the first turn.
- Every user (except one) initiates the follow behaviour of the robot after the greeting.

- Half of the users experienced problems with the follow function.
- All users, except two, succeeded with a show mission before experiencing some type of miscommunication.

If we look at the characteristics for the cases for which there are more than a few users, i.e. *Greeted* (all), *Problem/No problem with Follow* (half/half) and *Show worked* (most users). Using this *dynamic* information we could initially divide users into the following groups:

- Initial Greeting
- Could use follow
- Could use show

The question is what we could do in terms of design of a the dialogue. First of all, greeting allows the system to choose the next move, for instance, to suggest a set of tasks that are available. And, given the amount of users that greets the robot, any deviance from the initial-greeting pattern could be interpreted as a possible case of miscommunication.

1.2 Manipulators or Pointers?

In figure 2 the use of ergotic gestures, e.g., moving objects about, while preparing to show objects to the robot and the use of deictic gestures, i.e., pointing to objects and locations.

When looking at the sequences of interaction it appears that a preparation, i.e., *Manipulation precedes* Pointing, is a typical sequence. However, when looking at the data it appears that some of the users are manipulating their environment less than others. Hence, we could distinguish between users by using the ratio between manipulation, referring to one category as *manipulators* and the other one as *pointers*.

There are some problems with this approach, as there are obvious cases where there is no difference between the amount of manipulation and the amount of pointing gestures. Furthermore, if the user is pointing to *locations* these are not being manipulated at all.

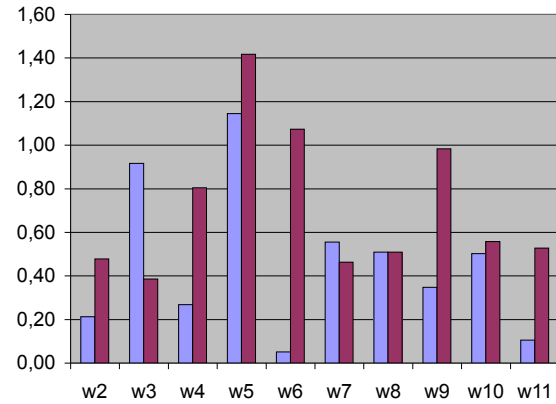


Figure 2: The figure shows pairs consisting of manipulations *per minute* (left bar) and referencing objects through pointing *per minute* (right bar)

2 Conclusion

Since we are using data from a study that only concerns the very first encounter between the robot and the user we should be careful with respect to the generality. However, we believe that the first experience with the robot is important for what happens next time the user encounters the robot. As any type of problem, breakdown or successful use of a new feature provides an opportunity for learning the boundaries for the system (Martinovski and Traum, 2003) we still believe it important to model this first encounter.

We have briefly sketched two different type of information that could be used to group users in the very first robot encounter.

- Greeting, first actions and first problem encountered,
- Distribution of deictic and ergotic (manipulative) gesture.

The first type of information can be accounted for using, e.g., an information-state approach used in dialogue systems. The second type, provides a challenge for a sensor based system. Thus in order to encounter for this type of activities we need to detect both pointing gestures and ergotic gestures. It is therefore necessary to allow for any component that maintains a user model to be able to receive or get information from other system components and should be considered in system design.

Greeting	Greeting	Greeting	Greeting	Greeting	Greeting	Greeting	Greeting	Greeting	Greeting
Follow	Follow	Follow	Follow	Follow	Follow	Social	Follow	Follow	Follow
Follow failure	Show	Follow failure	Show	Show	Show	Mismatch	Show	Follow	Show
	Follow	Follow	Follow failure	Follow	Show	Follow	Follow failure	Follow	Follow failure
	Show		Follow	Show	Show repair		Follow	Show repair	Follow
	Follow			Robot malf.				Show	
	Follow failure								

Figure 1: The initial part of the user sessions. First interactions and the first actions leading up to a miscommunication or problematic event.

Since this work is explorative we will turn our future interest towards behavioural features of interaction that could be used to tie into other frameworks, for instance, the notion of engagement (Sidner et al., 2005; Sidner et al., 2004). And also, with respect to Fischer’s notion about the *player/non-player*, the notion of *presence as immersion* (Lombard and Ditton, 1997) and *willingness to suspend disbelief* (Lombard and Ditton, 1997).

References

- Kerstin Fischer. 2006. The role of users’ preconceptions in talking to computers and robots. In *Proceedings of the Workshop on ‘How People Talk to Computers, Robots, and Other Artificial Communication Partners’*, pages 112–130, Delmenhorst, Germany, April 21–23. Hansewissenschaftskolleg.
- Anders Green, Helge Hüttenrauch, and Kerstin Severinsson Eklundh. 2004. 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*, pages 575–580, 20–22 Sept.
- Anders Green, Helge Hüttenrauch, Elin Anna Topp, and Kerstin Severinsson Eklundh. 2006. Developing a Contextualized Multimodal Corpus for Human-Robot Interaction. In *Proceedings of the Fifth international conference on Language Resources and Evaluation LREC2006*.
- Matthew Lombard and Theresa Ditton. 1997. At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication*, 3(2), September. online journal.
- Bilyana Martinovski and David Traum. 2003. The Error Is the Clue: Breakdown in Human-Machine Inter-

action. In *proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16.

- Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *IUI’04: Proceedings of the 9th international conference on Intelligent User Interfaces*, pages 78–84, New York, NY, USA. ACM Press.
- Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166:140–164.

The implementation of dialog adaptation

1 Realization of Adaptation.....	1
1.1 Overall structure and task devision.....	2
1.1.1 InteractionManager.....	2
1.1.2 User	2
1.1.3 AdaptationSupervisor.....	2
1.1.3.1 AdaptStrategyProvider.....	2
1.1.3.2 AdaptStrategyScheduler.....	2
1.1.3.3 AdaptStrategyExecutor.....	3
1.1.4 DiscourseObserver.....	3
1.1.5 Threshold.....	3
1.2 Triggers.....	4
1.3 Strategies.....	5
1.3.1 Level of strategies.....	5
1.3.2 Strategy scheduling.....	5
1.4 Evaluating the performance of the user and the robot.....	6
1.4.1 General info.....	6
1.4.2 In-interaction performance measurement for robot and user.....	6
1.4.2.1 Performance measurement	6
1.4.2.2 Firing trigger using threshold:.....	7
1.4.3 Overall performance measurement for robot and user.....	7
1.4.3.1 Performance measurement.....	7
1.4.3.2 Contextual consideration.....	8
1.5 Strategy updates.....	8
1.5.1 Initiation of the relevant values.....	8
1.5.2 Two types of strategy updates.....	8
1.5.3 User behavior observation.....	9
1.5.4 How the updating works.....	9
1.5.5 Currently implemented updates.....	10
1.5.5.1 Operations (class InteractionConstants).....	10
1.5.5.2 Initial settings for novice user (class Threshold).....	11
1.5.5.3 trigger-adoption catalog (adptStratDef.xml).....	11
1.5.5.4 trigger-update catalog (see 1.5.4How the updating works).....	11
1.5.6 Notes	11
1.6 Modification of the existing adaptation abilities.....	11
1.6.1 Adding new abilities.....	11
1.6.1.1 Adding new strategy for one existing trigger.....	11
2 Attachment.....	12
2.1 Definition of trigger-strategies: AdaptStratDef.xml.....	12
2.2 User session data: allUserData.xml.....	13

1 Realization of Adaptation

The biggest problem / danger in the implementation is the quantization of all the thresholds.

1.1 Overall structure and task devision

1.1.1 InteractionManager

the interface between the dialoging / roboting and the adaptation classes. Classes of the package dialoging and roboting call the adaptation classes via the InteractionManager.

1.1.2 User

models all the information that models the user:

static feature: String name, int id (can be variable from session to session), boolean isCP

dynamic features: interactionFrequency, interactionType (player vs. non-player) and helpSeekingType (helpAppreciater vs. helpRejecter).

1.1.3 AdaptationSupervisor

the supervisor of the adaptation process. Receives messages on trigger fire from the InteractionManager and organizes the adaptation process as follows:

1. Call [AdaptStrategyProvider](#) to get the appropriate strategies for the current trigger.
2. Call [AdaptStrategyScheduler](#) to sort the vector of strategies that this class receives from the [AdaptStrategyProvider](#) and schedule the order that the strategies should be executed.
3. Call [AdaptStrategyExecutor](#) to execute the strategies.

1.1.3.1 AdaptStrategyProvider

This class contains hard-coded policy for “*what-strategy-should-be-activated-for-the-user-given-the-trigger*”. It is also responsible for updating strategies when user behaviors are observed that affect the strategy policy in general. (for strategy policy updates, see 1.5.4How the updating works) Most important variables and methods are:

<i>AdaptStrategyProvider</i>	
Variables	<pre>private Vector tStrategies; private Vector sStrategies; private String sStratGroup; private String tStratGroup; private int tInitiativeFrequency; private int sInitiativeFrequency;</pre>
Methods	<pre>getStrategiesForTrigger() updateGroupForCurrentUser() SetInitialStrategiesForNoviceUser() setInitialStrategiesForKnownUser()</pre>

1.1.3.2 AdaptStrategyScheduler

The scheduling is especially important if there are still strategies to be executed left out in the vector [sortedStrategies](#) as a new trigger is fired. For more detail on the handling see 1.3.2Strategy scheduling:

<i>AdaptStrategyScheduler</i>	
Variables	<pre>private static Vector sortedStrategies; private int indexOfStratBeingUsed; private boolean moreStrategiesAvailable;</pre>
Methods	<pre>getStrategiesForTrigger() getNextStrategy() expandSortedStrategiesWith()</pre>

1.1.3.3 *AdaptStrategyExecutor*

This class directly calls the [DiscourseOperater](#) to create the appropriate exchanges. In case of normal exchanges that expect user reply, the crucial handling of user answer is done in the [DialogManager](#) and with the configuration file [meaningEffectMatch.xml](#). (see)

1.1.4 DiscourseObserver

This class is responsible for checking out each Exchange that is put into the discourse to find out:

1. whether a trigger for adopting a strategy can be fired
2. whether a trigger for updating the strategy policy can be fired
3. count how many errors the user and the robot create during the interaction
4. count how many meaningful exchanges are proposed by the user. (meaningful exchanges are exchanges the presentation of which was initiated by the user. This policy excludes the influence of system initiated PwoA and confirmation questions.)

For more info about this class, see 1.4Evaluating the performance of the user and the robot

<i>AdaptStrategyProvider</i>	
Variables	<pre>private static int user_Problems_DS_Excluded private static int user_Problems_DS_CurImpossible private static int user_Problems_Total private static int robot_Problems_DS private static int robot_Problems_Total private static int meaningful_Exchanges_DS private static int meaningful_Exchanges_Total</pre>
Methods	<pre>AnalyseExchange() analyseDiscourseSegment()</pre>

1.1.5 Threshold

This class maintains a list of constants that serve as thresholds and default values concerning the adaptation.

<i>AdaptStrategyProvider</i>	
Constants	<pre>ds_Perform_Remark_Threshold (t-initiative frequency) user_Ds_Perform_Bad_Threshold (s-initiative frequency)</pre>

<i>AdaptStrategyProvider</i>	
Methods	decreaseSInitiativeFrequency() decreaseTInitiativeFrequency()

1.2 Triggers

There are two types of triggers:

- strategy triggers: these triggers trigger direct execution of strategies. They are activated immediately after the event happens, e.g., after it is recognized that user gives a Cmd_excluded. They can be activated in [RobotingManager](#), [DialogManager](#) and [DiscourseObserver](#).
- update triggers: these triggers trigger the update of the strategy policy, and thus user profile. They are activated in the class [DiscourseObserver.analyseExchange\(\)](#)¹ every time when an exchange is inserted into the discourse which is done in the class Discourse. In case a trigger is fired, the [DiscourseObserver](#) calls the [InteractionManager.analyseUserBehavior\(\)](#).

The corresponding constants are specified in InteractionConstants.

<i>Type</i>	<i>Trigger</i>	<i>Type</i>	<i>Location</i>	<i>Remark</i>
Strategy trigger	TRIGGER_INTER_PRE_START	s	InteractionManager	Initially activated in RobotingManager
	TRIGGER_INTER_START	t	RobotingManager	This strategy consists several steps and are continued (in the DialogManager) if the user is cooperative.
	TRIGGER_CMD_CUR_IMPOSSIBLE	t	DialogManager	
	TRIGGER_CMD_EXCLUDED	t	DialogManager	
	TRIGGER_USER_BAD_PERFORM	s	DiscourseObserver	
	TRIGGER_USER_GOOD_PERFORM	s	DiscourseObserver	
	TRIGGER_ROBOT_BAD_PERFORM	s	DiscourseObserver	
	TRIGGER_INTER_END	s	DiscourseObserver	
Update trigger	TRIGGER_USER_DELETE_T		DiscourseObserver	
	TRIGGER_USER_ERROR_AFTER_T		DiscourseObserver	
	TRIGGER_USER_ERROR_AFTER_S		DiscourseObserver	
	TRIGGER_USER_SAYS_SORRY		DialogManager	

¹ As to the firing of trigger, the difference between the class DiscourseObserver and the DialogManager is, that the DiscourseObserver only does something after the processing of one exchange is already completed and it is added into the discourse, while the DialogManager can fire triggers before the processing of one exchange is completed.

1.3 Strategies

1.3.1 Level of strategies

The strategies are defined in the file of [adaptStratDef.xml](#). They can be seen as following the hierarchy:

- Level 1: Type (t or s-strategies)
- Level 2: group (max, mid, or min)
- Level 3: trigger
- Level 4: priority

For level 4 (priority), even if the first 3 levels are all identical and the content of the tag also seems to be identical, the plain text entries in the [outputTemplate.xml](#) are different for different priority. In fact, the priority is the content that is Intention that is passed to Checker for searching plain text. For more of the implementation, see [interactiig.AdaptStratExecuter.handleInterStart\(\)](#).

1.3.2 Strategy scheduling

In case of there are still strategies left out for the last trigger as a new trigger is fired, the corresponding strategies for this new trigger is added to the end of the Vector [sortedStrategies](#) that the [AdaptStratScheduler](#) maintains and are executed after all the strategies for the last trigger was all executed.

[AdaptationSupervisor.handleTrigger\(\)](#):

```
/*1. call the provider to provide the appropriate strategies for the new trigger.
Vector strategiesToBeHandled =
    provider.getStrategiesForTrigger(trigger, currentUser);
if(strategiesToBeHandled != null){
    /*2.1. if there are other strategies left out for the last trigger, expand the vector of
    strategy that the scheduler maintains. The current strategy is executed after all the last
    trigger's strategies are executed. */
    if(scheduler.moreStrategiesAreAvailable() &&
    //if the existing strategies doesn not have the same trigger as the new one
    !trigger.equals(((AdaptStrategy)(scheduler.getSortedStrategies().get(0)))
    .getTrigger())){
        strategiesToBeHandled =
            scheduler.expandSortedStrategiesWith(strategiesToBeHandled);
        return;
    }
    /*2.2. if all the strategies for the last trigger are already executed, fill the scheduler
    with the current strategies. */
    else
        scheduler.setAndSortCurrentStrategies(strategiesToBeHandled);
```

```

/*3. execute the strategies */
adoptStrategy(strategiesToBeHandled, currentEx,
              originalIntention, isContinuedAdaptation);

```

1.4 Evaluating the performance of the user and the robot

1.4.1 General info

The class `interacting.DiscourseObserver` is set up to centrally resolve this issue.

1. There are two types of performance that are measured:
 - In-interaction performance: the performance of the *robot* within one Discourse Segment (although the evaluation is not done at the end of the Segment, for more, see 1.4.2 In-interaction performance measurement for robot and user). This performance evaluation is associated with the trigger `TRIGGER_USER_BAD_PERFORM`, `TRIGGER_USER_GOOD_PERFORM` and `TRIGGER_ROBOT_BAD_PERFORM`.
 - Overall performance: the performance of *both* dialog partners that is measured at the end of the whole session. This performance evaluation is associated with the trigger `TRIGGER_INTER_END`. The decision making process of whether this triggers should be fired also involves the contextual consideration (method `discourseContextAllowsSessionSummary`).
2. The errors of the users and robot that are measured are:
 - The errors of the user include: `Cmd_cur_impossible` and `Cmd_excluded`.
 - The errors of the robot include: `Clar_nonUnderst_verbal`, `Clar_partialUnderst` and `Clar_nonUnderst_nonverbal`.
3. The methods of this class is called in the `dialoging.model.Discourse.addExchange()`:
 - each time when a new exchange is added into the discourse, this exchange will be analyzed.
 - each time when a new discourse segment is created, the last segment will be analyzed.

1.4.2 In-interaction performance measurement for robot and user

“In-interaction performance” is the performance of the robot and the user *after three meaningful exchanges*. In the class of `DiscourseObserver`, the variable concerning the problems that user and robot created are increased every time they occur.

1.4.2.1 Performance measurement

1. The performance is measured with parameters (a) the number of user/robot problems and (b) the number of meaningful exchanges so far, with the following formula:

$$P_{r/u} = E_{\text{problem}} / E_{\text{total}}$$
 where the E_{problem} means is the number of exchanges that indicates interaction problems (both).
2. User problems are measured using (`DiscourseObserver.analyseExchange()`):


```

if(currentIntention.getType().equals(Constants.INTENT_CMD_EXCLUDED) ||
   currentIntention.getType().equals(Constants.INTENT_CMD_CUR_IMPOSSIBLE))
    user_Problems_DS++;

```
3. robot problems are measured using (`DiscourseObserver.analyseExchange()`):

```

if(currentEx.getIntention().getType().indexOf(Constants.PREFIX_PROBLEM)>= 0)
    robot_Problems_DS++;

```

The robot's problems are perceptual problems including: `Clar_nonUnderst_verbal`, `Clar_partialUnderst` and `Clar_nonUnderst_nonverbal`.

1.4.2.2 *Firing trigger using threshold:*

4. The calculated $P_{r/u}$ is then compared with a threshold that decides whether the performance is good / bad.
 - In the actual implementation, the threshold `Threshold.ds_Perform_Remark_Threshold` (for doing remark on the performance of both user and the robot as s-initiative) or `Threshold.user_Ds_Perform_Bad_Threshold` (for providing task-related help as t-initiative) is multiplied with the $EX_{problem}$, if the result is then greater than or equal to the EX_{total} , then the threshold is exceeded.
 - Based on this calculation, if the `Threshold.ds_Perform_Remark_Threshold` is set to 3, then it means that the $EX_{problem}$ should be at least 1/3 of the EX_{total} to fire the trigger. Thus, the higher this value is, the more frequent should the robot make the remark. If the value is 0, the robot should completely stop doing remark.
 - During the adaptation, every time when the user ignores or reject such comments from the robot, the value of these two thresholds is decreased. Once they arrives 0, then the user's `helpSeekingType` and `interactionType` will change to the passive one.

1.4.3 Overall performance measurement for robot and user

1.4.3.1 *Performance measurement*

1. The overall performance is measured during the interaction following the principles described in the last section. The variables `robot_problem_total` and `user_problem_total` are increased with 1 every time when they create an error.
2. After user initiates the exchange “Parting”, the overall performance will be measured based on the problems that have been created by the dialog partners (method `getTroubleMakerOfTheSession()`):

```

if(robot_problems_total - user_problem_total >= Threshold_1)
    trouble maker = robot;
else if(user_problem_total - robot_problems_total >= Threshold_1)
    trouble maker = user;
else if((robot_problems_total + user_problem_total) * Threshold_2 >=
meaningful_Exchanges_total)
    trouble maker = both;
else
    trouble maker = none;

```

where `Threshold_1 = Threshold.ROBOT_USER_PERFORM_DIFF_THRESHOLD`;

`Threshold_2 = Threshold.BOTH_BAD_TOTAL_PERFORM_THRESHOLD`

1.4.3.2 Contextual consideration

The trigger of TRIGGER_INTER_END can only be fired if

- user initiates a Parting exchange (by saying good-bye)
- current (Parting) exchange does not have son-exchanges with support relations (because otherwise the Parting exchange initiated by the user created problem (e.g., Cmd_cur_impossible) so that the robot initiates support exchanges).

This consideration is done in the method: `discourseContextAllowsSessionSummary()`

1.5 Strategy updates

1.5.1 Initiation of the relevant values

1. The current user is initially set in `RobotingManager.processPTAInfo()`. Then after the `dlgInInteraction` is set to true in `CommWatcher.processInput`, a `AdaptationSupervisor` is assigned to the current user to take care of the adaptation.
2. The relevant adaptation values are:
 - t-StrategyGroup
 - s-StrategyGroup
 - t-StrategyFrequency (associated with `Threshold.user_Ds_Perform_Bad_Threshold`)
 - s-strategyFrequency (associated with `Threshold.ds_Perform_Remark_Threshold`)
3. For an experienced user, these values will be initiated with the ones specified in the `allUserData.xml` (which means the values from the last session); for a novice user, the values are set to the default ones in the class `AdaptStrategyProvider`. For all the initial settings in the current implementation, see 1.5.5Currently implemented updates.

1.5.2 Two types of strategy updates

There are two types of updating:

1. updating the strategy families: the fundamental content of the strategies are changed because the strategy groups (mid, min, max) are changed.
2. Updating the thresholds: The content of the strategy is not changed, but only its frequency, which is implemented as the Threshold for firing the corresponding triggers. If the frequency is reduced to 0, this special strategy is completely excluded. Once this happens, an strategy family update will follow. This is an idea of “changing quantity leads to changes of quality”. Such updates are applicable for variable strategies (as indicated by the attribute “variable” in the `adaptStratDef.xml`). More specifically,
 - such variable t-strategy is `InteractionConstants.STRAT_CONTENT_PERSONALIZED_ABILITY_EXPLAIN` with the trigger `InteractionConstants.TRIGGER_USER_BAD_PERFORM`.
 - And such s-strategies are `STRAT_CONTENT_PRAISE` and `STRAT_CONTENT_APOLOGY` for trigger `TRIGGER_ROBOT_BAD_PERFORM` and `TRIGGER_USER_GOOD_PERFORM`.

These updates are executed in the `interacting.AdaptationSupervisor.analyseUserBehavior()`.

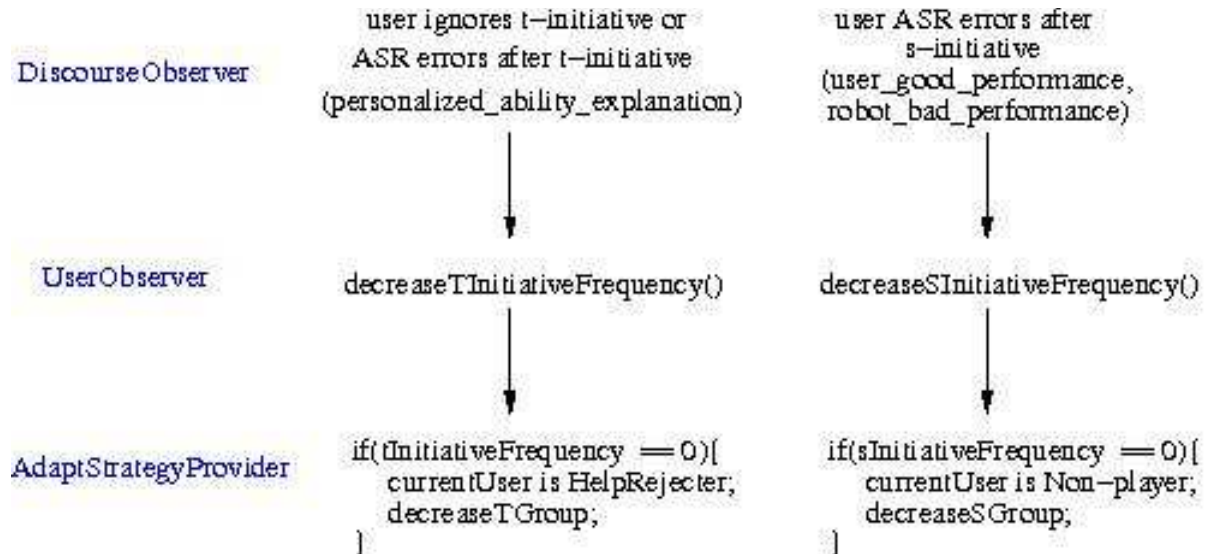
1.5.3 User behavior observation

To update the adaptation policy, the system observes user's reaction to its own initiatives for the t- and s-strategies. This information is mainly delivered by the `DiscourseObserver`. Relevant user behaviors for such updates are:

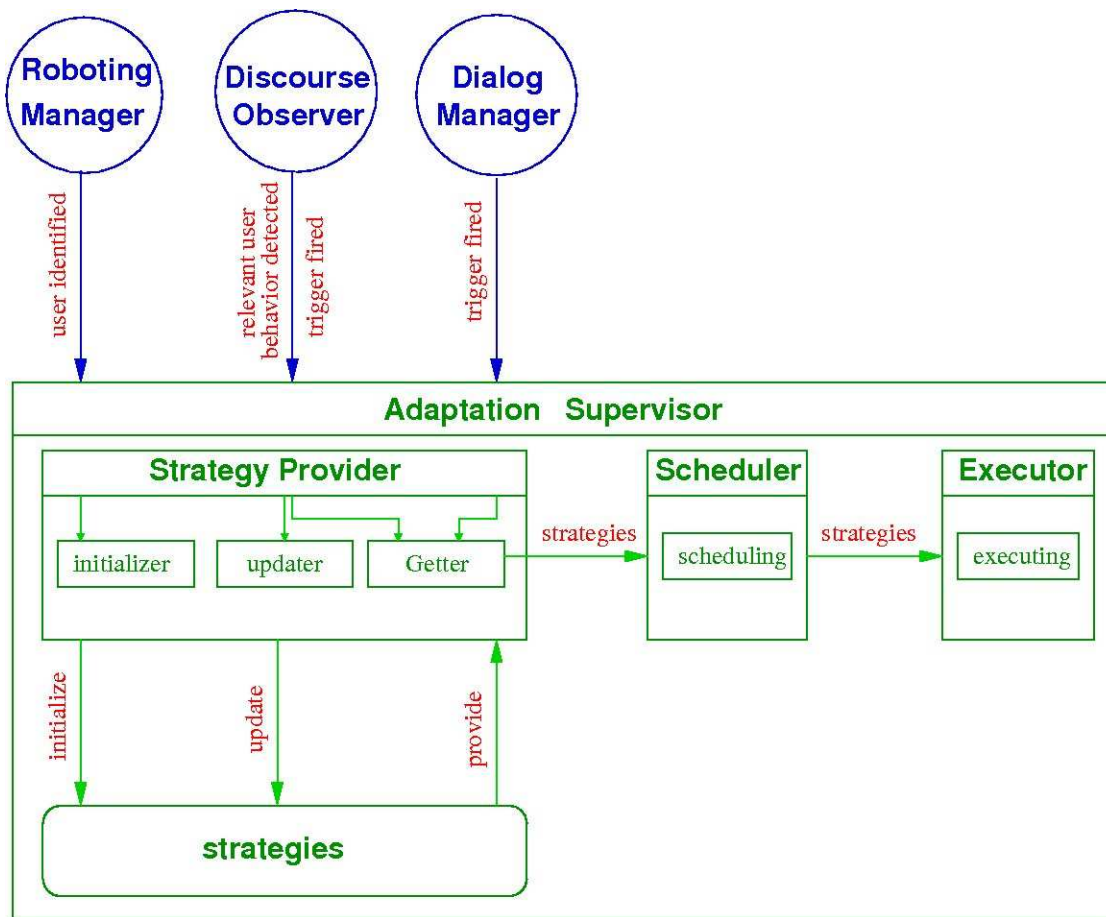
- whether user ignores system's t-initiative
- speech recognition errors after system's t-initiative
- speech recognition errors after system's s-initiative
- user says “thank you” or “sorry”.
(`if(intentionType.indexOf(Constants.INTENT_JUDGE_COMM) >= 0)`)

For currently implemented if-else-strategy-update catalog, see 1.5.5Currently implemented updates

1.5.4 How the updating works



A more general graph:



1.5.5 Currently implemented updates

1.5.5.1 Operations (class *InteractionConstants*)

```

public static String META_ADAPT = "____ADAPT:";
/*trigger the update of the strategy-group or frequency*/
public static String META_TRIGGER_UPDATE = META_ADAPT + "TRIGGER_UPDATE____";
/*trigger the adoption of a strategy*/
public static String META_TRIGGER_ADOPTION = META_ADAPT + "TRIGGER_ADOPTION____";
/*continue the adoption of a strategy*/
public static String META_CONTINUE_ADOPTION = META_ADAPT
+"CONTINUE_ADOPTION____";
/*update a strategy*/
public static String META_UPDATE = META_ADAPT + "UPDATE____";
/*adoption of a strategy*/
public static String META_ADOPTION = META_ADAPT + "ADOPTION____";

```

1.5.5.2 Initial settings for novice user (class [Threshold](#))

```
DEFAULT_EXPLEVEL = InteractionConstants.EXPLEVEL_LOW;  
DEFAULT_INTERTYPE = InteractionConstants.INTERTYPE_PLAYER;  
DEFAULT_HELPSEEKTYPE = InteractionConstants.HELPSEEKING_TYPE_APPRE;  
DEFAULT_S_INITIATIVE_FREQUENCY = 3;  
DEFAULT_T_INITIATIVE_FREQUENCY = 3;
```

1.5.5.3 trigger-adoption catalog ([adptStratDef.xml](#))

See the file [adptStratDef.xml](#) and compare 1.2Triggers

1.5.5.4 trigger-update catalog (see 1.5.4How the updating works)

<i>Trigger</i>	<i>Effect</i>
TRIGGER_USER_DELETE_T	Decrease t-initiative frequency
TRIGGER_USER_ERROR_AFTER_T	Decrease t-initiative frequency
TRIGGER_USER_ERROR_AFTER_S	Decrease s-initiative frequency
TRIGGER_USER_SAYS_SORRY	Increase s-initiative frequency

1.5.6 Notes

The definition of the self performance remarks:

- each time speech recognition errors occurs immediately after *each* system's s-initiatives, the threshold for performance remarks is decreased with 1.
- This means, the performance remarks can be completely shut down even before the system ever evaluated one discourse segment.
- If several speech recognition errors occurs after one s-initiative, the mother exchanges of these corresponding exchanges will be set to the initial s-initiative. For one t- or s-initiative, the Threshold only decreases once.

1.6 Modification of the existing adaptation abilities

1.6.1 Adding new abilities

1.6.1.1 Adding new strategy for one existing trigger

1. Adding the definition of a new strategy (mainly its attribute *content* of a tag, but if there are

several content matching one strategy, also the attribute *priority*) in the configuration file of `adaptStratDef.xml`.

2. Adding the corresponding constants in the class of `dialoging.Constants` and `interacting.InteractionConstants`.
3. Adding the response plain text in `outputTemplate.xml`.
4. Adding its discourse segmentation purpose in the `taskCollection.xml`.
5. Updating the class of `interacting.AdaptStratExecutor` with a part of an existing method that executes this strategy.
6. If this strategy requires user reply, i.e., in the `AdaptStratExecutor` a normal exchange is created instead of a `PwoA`, then also adding the requirements for an user acceptance in the file of `meaningEffectMatch.xml`.

2 Attachment

2.1 Definition of trigger-strategies: *AdaptStratDef.xml*

```
<adaptStratDef>
  <strategies type = "t">
    <!--<strategy group = "min" variable = "false" trigger =
"Question_abilities" priority = "1">adt_answer</strategy>-->
    <strategy group = "max" variable = "false" trigger =
"Cmd_cur_impossible" priority = "1">adt_confirmation_question</strategy>
    <strategy group = "mid" variable = "false" trigger =
"Cmd_cur_impossible" priority = "2">adt_propose_new_task</strategy>
    <strategy group = "mid" variable = "false" trigger = "Cmd_excluded"
priority = "1">adt_propose_new_task</strategy>
    <strategy group = "max" variable = "false" trigger =
"interaction_start" priority = "1">adt_name_explanation</strategy>

    <!--the first set of ability explanation, to be identified using
priority -->
    <strategy group = "max" variable = "false" trigger =
"interaction_start" priority = "2">adt_propose_ability_explanation</strategy>
    <strategy group = "max" variable = "false" trigger =
"interaction_start" priority = "3">adt_self_ability_explanation</strategy>
    <!--the second set of ability explanation, to be identified using
priority -->
    <strategy group = "max" variable = "false" trigger =
"interaction_start" priority = "4">adt_propose_ability_explanation</strategy>
    <strategy group = "max" variable = "false" trigger =
"interaction_start" priority = "5">adt_self_ability_explanation</strategy>
    <!--the third set of ability explanation, to be identified using
priority -->
    <strategy group = "max" variable = "false" trigger =
"interaction_start" priority = "6">adt_propose_ability_explanation</strategy>
    <strategy group = "max" variable = "false" trigger =
"interaction_start" priority = "7">adt_self_ability_explanation</strategy>

    <strategy group = "max" variable = "false" trigger =
```

```

"interaction_start" priority = "8">adt_propose_test_run</strategy>
    <strategy group = "mid" variable = "true" trigger =
"user_bad_performance" priority =
"1">adt_personalized_ability_explanation</strategy> <!--apology if his performance
is good-->
    </strategies>

    <strategies type = "s">
        <strategy group = "mid" variable = "false" trigger =
"interaction_pre_start" priority = "1">ads_execute_initiative_greeting</strategy>

        <strategy group = "min" variable = "false" trigger = "interaction_start"
priority = "1">ads_show_happiness_for_user</strategy>
        <strategy group = "min" variable = "false" trigger = "interaction_end"
priority = "1">ads_summarize_interaction</strategy>
        <strategy group = "mid" variable = "true" trigger =
"user_good_performance" priority = "2">ads_praise</strategy> <!--praise the user if
his performance is good-->
        <strategy group = "mid" variable = "true" trigger =
"robot_bad_performance" priority = "1">ads_apology</strategy> <!--apology if the
performance of the robot is bad-->
    </strategies>

</adaptStratDef>

```

2.2 User session data: allUserData.xml

<!--In the following, the information on two users who have interacted with the system is recorded.-->

```

<?xml version="1.0" encoding="UTF-8"?>
<users>
  <user>
    <name>sonja</name>
    <id>19</id>
    <interactionFrequency>2</interactionFrequency>
    <interactionType>player</interactionType>
    <helpSeekingType>rejecter</helpSeekingType>
    <lastSessionSettings>
      <tStrategyGroup>max</tStrategyGroup>
      <sStrategyGroup>max</sStrategyGroup>
      <tStrategyFrequency>0</tStrategyFrequency>
      <sStrategyFrequency>2</sStrategyFrequency>
    </lastSessionSettings>
  </user>
  <user>
    <name>Tom</name>
    <id>256</id>
    <interactionFrequency>1</interactionFrequency>
    <interactionType>player</interactionType>
    <helpSeekingType>appreciater</helpSeekingType>
    <lastSessionSettings>
      <tStrategyGroup>mid</tStrategyGroup>
      <sStrategyGroup>mid</sStrategyGroup>
      <tStrategyFrequency>4</tStrategyFrequency>
      <sStrategyFrequency>3</sStrategyFrequency>
    </lastSessionSettings>
  </user>
</users>

```

```
</user>  
</users>
```

Integrating a Human Oriented Map Representation for Exploring new Areas in a Team of Human and Robot

Thorsten Spexard, Shuyin Li,
Britta Wrede, Marc Hanheide, Gerhard Sagerer
Applied Computer Science
Bielefeld University
D-33501 Bielefeld, Germany
{tspexard, shuyinli, bwrede, mhanheid, sagerer}
@techfak.uni-bielefeld.de

Elin A. Topp, Helge Hüttenrauch
School of Computer Science and Communication
Royal Institute of Technology (KTH)
SE-100 44 Stockholm, Sweden
{topp, hehu}@csc.kth.se

Abstract— A very important goal in current research of social robots is the cooperation in a team of human and robot. Especially in real world surroundings like home environments where personal service robots are instructed by several people, different challenging tasks have to be performed. One of these tasks is the robust spatial orientation in an unknown surrounding. In this paper we present the conceptual integration of topological localization using a map supporting attention hierarchies as derived from user studies. We will demonstrate the benefits of combining this localization approach with a powerful spoken dialog system. It will be used not only for labeling but, taking the special map representation into account, for learning and tour guiding in a human-like interaction between a robot and its owner. These enhanced interaction capabilities that will provide an extended human robot cooperation are crucial for the aim of developing a new generation of robots that will be accepted not only as service robots but also as robot companions in our daily life.

I. INTRODUCTION

With the progress in research and product development humans and robots get more and more close to each other. Specialized service robots (e.g., vacuum cleaners) have already entered people's homes and are used on a regular basis. We suggest a more general view on service robots by working with the idea of a robot companion, a service robot that is assigned to a particular working environment and has a limited number of users. This is based on the idea of the "Cogniron – The cognitive robot companion" [3] project which builds the framework to the work presented here.

A "personal" (companion) service robot is designed to fulfill a number of tasks in a domestic setting that require indoor localization and navigation and thus very good knowledge of the environment. This knowledge includes also the locations of particular items (e.g., the refrigerator, in case the task set comprises fetch-and-carry of cooled drinks) or the particular looks of smaller objects that might have to be manipulated (e.g., a "personal cup").

In previous work we already suggested the framework of "Human Augmented Mapping" to obtain an initial shared environment representation for a service robot and its user by integrating the user into a robotic mapping process in an interactive scenario (the "guided home tour") [21], [22],

[24]. This framework allows us to augment a robotic map with information given by the user, while communication is facilitated with the help of an adequate generic environment model. This model needs to provide appropriate concepts (or categories) to be filled with the information given by the user. Since human encoded knowledge of spatial relations and layouts is most likely represented in a partial hierarchy [17], we suggest to use a hierarchical approach to model the link between robotic mapping and human adequate communication. Other hierarchical approaches to environment representations have been used to model exploration strategies [13] or wayfinding routines for robots [11], mostly in the context of autonomous systems.

This paper presents our effort of joining a robotic mapping approach through a hierarchical environment model and an attentional hierarchy with a multi-modal dialogue and interaction control system. The hierarchy is motivated from concepts investigated in Cognitive Science and anecdotal evidence from user studies. We show the link from the concepts and the handling of different situations that can occur during a "guided tour" scenario to the dialogue and propose a communicational framework for the actual integration into a working system on software level. The environment modeling and the robotic mapping subsystem are thus used to provide the overall system with the ability to localize in a (for humans) topological sense and thus communicate its whereabouts appropriately.

A. Related work

Our aim is to integrate several areas of robotics and human-robot interaction into one conceptual framework to investigate a specific scenario, the "guided home tour". It is thus natural that the background for the work has to be seen in all those areas for themselves as also in other approaches to integrated and / or interactive frameworks. Strongly related to the work presented here, but conducted under a different focus is the work of Kruijff *et al.* [12]. Other work can mostly be related to parts of the integrating concept, e.g., in terms of the hierarchical approach to the environment model, but without the interaction [7], or integrating mapping and interactively given information, but without the conceptual

modeling [4]. Other references, directly related to the particular parts of the system, are given in relation to the description of those.

B. Organization of the paper

The remainder of this paper is organized as follows. In section II we motivate our hierarchy from similar investigations in Cognitive Science and Robotics and give examples for how the hierarchy can manifest itself in human-robot interaction. Section III explains how this hierarchical model is implemented for the robotic mapping process and in section IV we present the dialogue model and its link to the environment model. With section V-B we propose a way of integrating all the necessary components into a communicational software framework on a robot. Conclusions are drawn in section VI where we also give some ideas for future work.

II. THE HIERARCHY

As mentioned before we use a hierarchy to describe the arbitrary indoor environments which we assume as the robot's workspace. So far we only incorporate three layers into this hierarchy, namely *rooms*, *areas* and *objects* (also termed in other publications as *regions*, *locations* and *objects* [21], [23]). A *room* is a somehow *delimited* part of an indoor environment. In most cases this corresponds directly to what common sense understands as “room”, but it could also be a cubicle or otherwise separated area. With an *area* we mean a specific but not spotlike part of a room where something of interest for the user and the robot is located. This could be, for example, the sofa or the TV-set. Areas can be used to describe a certain workspace for the service robot, as in “pick up my glasses from the table, please”, where the task for the robot means to go “somewhere close to the table” (into the *table area*) to determine from which position the glasses can be picked up. *Objects* are not represented in the robotic mapping hierarchy, but they are incorporated in the environment model to reflect the domestic situation. An *object* is a mobile item that can and should be manipulated by the robot, e.g., a cup, glasses, a bottle, or the remote control. For the recognition of objects an Object Attention System is integrated independently from the mapping process.

III. MAP REPRESENTATION IN LOCALIZATION

The robotic localization and mapping system used as one of the major parts for the integrated system relies on different, hierarchically organized components. Figure 1 gives a schematic overview of this mapping subsystem. Based on a combination of metric and topological methods the system is able to report its whereabouts in a human-comprehensive way using concepts and topological information, while exact navigation can still be enabled through an underlying metric mapping system. For this paper the most relevant level of this mapping hierarchy is the highest level, the environment model, as described in the previous section. A central question for the topological/conceptual localization the subsystem has to provide is how to map the robotic sensory data into

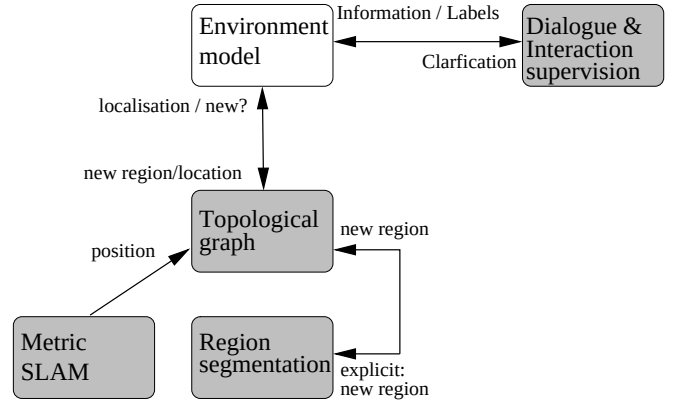


Fig. 1. Overview of the mapping subsystem

the representation based on concepts. In the following two approaches to answer this question are given.

A. Representing rooms

In recent work [23] we propose a laser range data feature based description of rooms (regions) that can be used for the segmentation of the environment into the necessary topological nodes (guided by the user's information to form the nodes corresponding to the spatial concepts). Additionally the feature based description allows to facilitate the localization in terms of rooms to a certain degree; in cases the underlying metric localization process becomes uncertain, the feature description can be used to generate hypotheses for localization in the topological sense. This approach is similar to a method for learning of places used by Mozos *et al.* [18].

The features used are the results of a statistical analysis of a laser range data set taken at a point specified by the user. Using a small number of features (the length and width of the data set in the directions of the principal component vectors as well as their ratio to each other and the covered area (mass) of the set) allows to keep a very concise description for relevant rooms.

B. Representing areas

One option to represent areas with the help of available sensor readings is to use a “view” representation of the important part of the room, taken from a pose from which the robot observed the area. A problem here is definitely that laser range data only can not always be sufficient to create a unique representation for each area. Thus, the pose (position and orientation) relative to the surrounding room is also taken into consideration to get an estimate of “where to go”, but the exact position can be reached or described with a certain margin. Another option to represent such a view is of course to use images of the particular observed scene and use a similarity measure of the appearance for the decision, where the area has its border [19]. So far we concentrate on the combination of laser range data and pose representation. In the rare case that both cues, laser range data and pose, fail the Dialog is used to ask a human in the robots surrounding

for assistance, learning information about its current position to avoid this situation next time.

IV. DIALOG SYSTEM

In general a dialog system is responsible for carrying out interactions with the user including transferring user commands to the robot control system and reporting task execution results to the user. During the conversation a dialog system should be able to regulate the initiative distribution, handle miscommunication, draw inferences between dialog participants' contributions and organize and maintain the discourse. To enable these abilities we implemented a powerful grounding-based dialog model [15] for BIRON.

Clark [2] proposed the notion of grounding: during a conversation the dialog participants need to coordinate their mental states based on their mutual understanding about the current intentions, goals and tasks. He termed this process as 'grounding'. Furthermore, a speaker can only be sure that her account (presentation) was fully understood if her dialog partner provides some evidence of understanding (acceptance), i.e., if the 'common ground' is available. Only then will the speaker be willing to proceed to another account. Our dialog system is based on this idea. We represent dialog participants' contributions as exchanges, i.e., pairs of contributions. They achieve 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*) and they can also have local effects on their previous exchanges. According to the execution results of the robot control system the dialog system formulates contributions for the robot. Each contribution of both the user and the robot is categorized in terms of its roles, 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 into a vector which records 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 ([25], [1]). The implemented system enables a mixed-initiative dialog style and a well-organized discourse maintaining mechanism. The first evaluation [16] with this Dialog System provides strong evidence for its flexibility and powerfulness. This system could facilitate smooth conversation concerning localization tasks in HRI by flexibly handling complex clarification questions based on the messages that will be sent by the Topological Localization.

V. USING LOCALIZATION AND DIALOG IN AN INTEGRATED SYSTEM

In the following we present more concrete use cases in which our integrated system will prove its benefit and performance. Therefore we will explore areas in the real world, which are unknown to a robot using the mobile demonstrator BIRON.

A. New Perspectives

This Exploration will be done in cooperation between robot and human, since the robot will not only be taught by a person in a natural way using a spoken dialog and deictic gestures for environment references, but will also be able to take the initiative by itself. This initiative ranges from simply asking for names of unknown classified areas to solving ambiguities in naming areas or objects. E.g., different people can name places differently or a certain kind of area in a room, a cupboard may occur in different rooms. For the first kind of ambiguity the robot could ask the user whether this is an alternative name for a location or if it made a mistake in classification. Thus the task *labeling* alone depends on three parameters that act in interplay:

- Lu: the label that the human user proposes for the current location. Users can name the same location differently for various reasons (e.g., by mistake, different users, at different time).
- Lc: robot's perception of the current location, i.e., what the robot thinks where it is. Robot can be wrong in its perception, it can also have registered a wrong label.
- Lm: The robot's knowledge of other locations.

The need for interaction arises when these three parameters do not correlate in a logical manner. For example, the user proposes the name A for the current location ($Lu = A$) while the robot "believes" that it is currently in an unlabeled location ($Lc = \text{new}$) and it already knows another location with the name A ($A \in Lm$). In this situation, possibly the user wants a second room to have the name A, or the robot's perception is wrong and it is actually in the known location A. However, if the robot does not know another location with the name A ($A \notin Lm$) then it is more likely that this is a normal operation of the user to label a location, which was previously unknown to the robot, with the name A. This would be a situation that hardly needs further negotiation between the user and the robot. To resolve such ambiguities the robot should be able to first initiate clarification questions and then relabeling or relocating itself according to the user's answer which requires a close cooperation between the Dialog System and the Topological Localization. Table I summarizes cases that occur most frequently in labeling tasks. For cases that require ambiguity resolution the relationship between the intention of the user (recognized by the Dialog System through clarification) and the corresponding operation of the Topological Localization is specified. This clarification request behavior is comparable to very young children as previously described in section II who expect only one label

Case Name	Parameter combination			User Intention recognized through clarification	Operation of Localization
	Lu	Lc	Lm		
initial labeling	A	new	$\emptyset$	–	label A
repeated labeling	A	A	A	–	–
double labeling	A	new	A	user wants a second room to be labled as A	add A
				user wants to correct Lc	relocate to A
correction	B	A	A	user wants to add a second name for the location	add B
				user wants to rename the current location	relabel A to B
relocate	B	A	A, B	–	relocate to B
lost	any	unknown	any	<i>(asks the user to take the robot back to an known location for re-orientation)</i>	

TABLE I
USE CASES FOR THE DIALOG SUPPORTED LOCALIZATION

for the same object or area and need additional explanation otherwise.

Besides ambiguities in the labeling process, ambiguities for areas or objects with the same name during data request should be solved. The topological map representation provides the information for the robot to discriminate a brush at the bathroom (for teeth cleaning) from a brush at the hallway for tidying up the room. So a fetch and carry task like "Please, fetch me the brush" might be solved differently according to the recent position of the robot, or at least the robot will suggest the according brush via the Dialog. For a robot already knowing its operating area a comfortable and more human-like way of guiding can be established. Human navigation often takes relations to objects into account, e.g., "The kitchen is to the right side of the stairs." instead of "Go 5 meters ahead, turn 90 degrees clockwise, go 2 meters ahead." In our scenario the robot can guide a interaction partner from one way point to another, with each way point being an entry on its topological map by following a person and politely asking her to turn around if she has gone into the wrong direction or guiding her to the next way point using the navigation.

B. Integration on a mobile demonstrator

Although the presented system is designed on arbitrary robots using a laser range finder and a dialog based user interface, we choose BIRON - The Bielefeld Robot Companion (see Fig. 2) as a demonstrator. It bases on the Pioneer2 PeopleBot from ActiveMedia. The robot is 1.45 meters tall and uses one laser range finder for the perception of the surrounding in front of it. Two color video cameras are mounted for general observation of the recent scene and for detailed focusing of persons and objects. At the top of the upper platform two microphones for stereo recording of sound sources are positioned. Additionally a touch screen display is mounted on the platform, too. The touch screen

both provides an intuitive feedback by an animated face representing the current robot state and could additionally provide context dependent menus for quick user interaction. Besides the visual feedback of the touch screen two speakers, recognizable as black circles are used for speech output initialized by the Dialog (see section IV). For computation two onboard PCs (Pentium III 500MHz, and Pentium III 850MHz with 256MB RAM each) are used in combination with three Laptops (Pentium Mobile, 1.7GHz, 1GB RAM) in a network. Due to the different calculation powers the onboard computers are in primary use for sensor reads and hardware control. The Dialog system including speech processing and the image processing are running separately each on one notebook. The third notebook will be used for

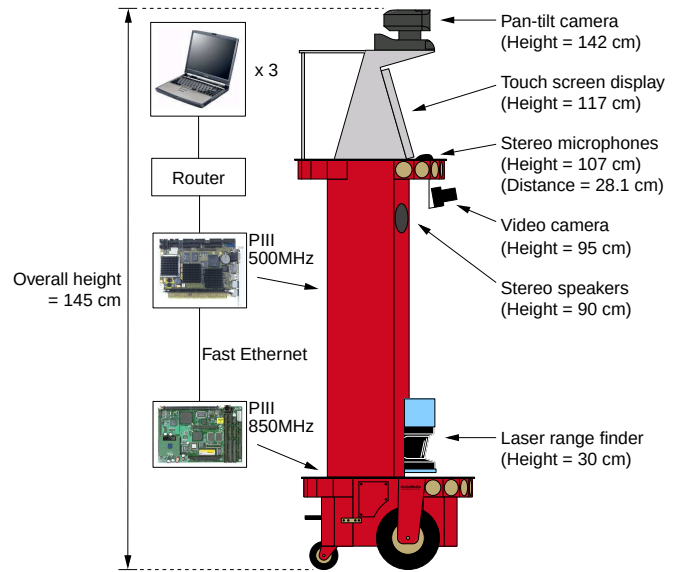


Fig. 2. Hardware composition of BIRON

the Topological Localization and Navigation to achieve a system running completely in real time.

For the realization of a real time capable system, operating in various home environments we chose a decentralized and modular hybrid system architecture [5] as first described by [8]. It consists of a pure reactive layer, containing modules with no or very poor planning or memory capabilities, a deliberative layer for modules with high planning or controlling capabilities, and an intermediate layer for the synchronization of the deliberative and the reactive layer (see Fig. 3). The synchronization task is solved by the so called Execution Supervisor which is in principle a finite state machine, reading the defined states and transitions from a configuration file written in XML. The Execution Supervisor uses the data output of the other modules (*Events*) to perform a transition from one robot state to another one. For each transition the Execution Supervisor can reconfigure the modules of the reactive layer (*Orders*) and informs the modules of the deliberative layer about the current system state (*Status*). Thus it can also forward information from one module to another but does not interpret it. In this way the data format can change and even new modules can be connected to the Execution Supervisor without changing one line of source code, but only editing the configuration file with the definition of the finite state machine. This architectural concept is sufficiently generic to be applied to a totally different anthropomorphic robot [20] using the same concept of deliberative and reactive elements.

For the successful integration process of the topological map representation and localization in our mobile robot, multiple modules as depicted in Fig. 3 are involved, to extend interaction capabilities in combination with the Dialog. As part of the deliberative layer the Dialog is considered as the mayor element of the control architecture. All tasks of the robot concerning the interaction with one selected person

are triggered by Events sent from the Dialog, based on user input. For the connection of the independently working modules of all three layers we use a simple to handle but powerful communication framework called *XML enabled Communication Framework* (XCF) [26]. As we concentrate on the expected benefits of this combination, the remaining modules will not be detailed here but are described in previous work [6], [14], [10], [9]. Using this modular approach for easily extending the abilities of the robot towards more cooperative interactions will enable us to investigate the upcoming questions concerning HRI in our daily surrounding. We are convinced that the dialog supported localization as a cooperative task, will raise the long term acceptance of a robot compared to systems accomplishing a calculated path, expecting the user to follow it without interaction.

VI. CONCLUSION & OUTLOOK

In this paper we presented a conceptual integration of a localization and mapping system with a spoken dialog system on a mobile demonstrator for a jointly exploration of unknown environment with a human. To achieve a human like representation of the surrounding user studies were conducted with the results Therefore the map representation was designed in a similar way using topologies and an attention hierarchy to report its whereabouts in a human-comprehensive way for the ordered representations of rooms, areas, and objects. The presented Dialog is used for transferring user commands, reporting results to the user and handling clarification questions. We described how these modules could be integrated into an existing software framework on the robot BIRON by a XML based data exchange. Thus aiming at a more human oriented way of a "guided home tour" compared to previous localization tasks in HRI. The benefit was demonstrated by different use cases, concerning the solving of ambiguities and error correction on the example of labeling and 'fetch and carry' tasks.

As a next step after realizing the described integration, the robot will extendedly be used in a real world flat, rent for user studies and long term evaluation of the described and upcoming abilities within the next years. Thus presenting convincing evaluation data and new results for the ongoing HRI research, soon.

VII. ACKNOWLEDGMENTS

This work has been supported by the European Union within the 'Cognitive Robot Companion' (COGNIRON) project (FP6-IST-002020) [3].

REFERENCES

- [1] 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.
- [2] H. H. Clark, editor. *Arenas of Language Use*. University of Chicago Press, 1992.
- [3] COGNIRON. The cognitive robot companion. <http://www.cogniron.org>.
- [4] A. Diosi, G. Taylor, and L. Kleeman. Interactive slam using laser and advanced sonar. In *Proceedings of the IEEE International Conference on Robotics and Automation*, Barcelona, Spain, Apr. 2005.

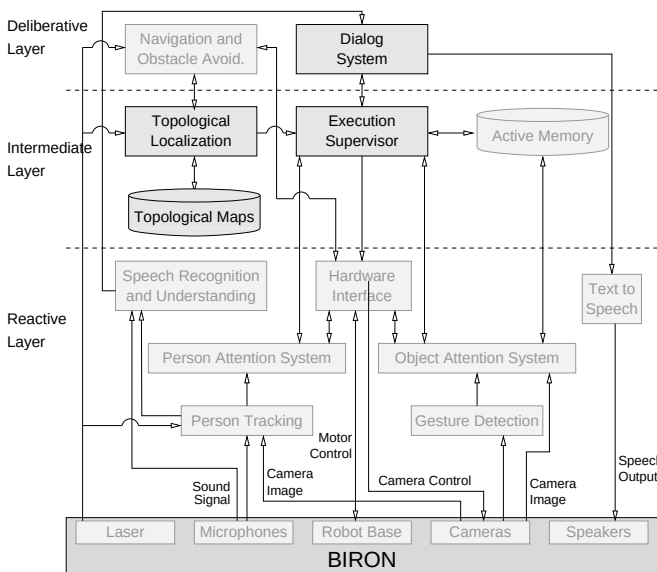


Fig. 3. Hybrid three layer architecture: The data of the deliberative and reactive modules is synchronized by the Execution Supervisor.

- [5] J. Fritsch, M. Kleinhagenbrock, A. Haasch, S. Wrede, and G. Sagerer. A flexible infrastructure for the development of a robot companion with extensible HRI-capabilities. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 3419–3425, Barcelona, Spain, April 2005.
- [6] J. Fritsch, M. Kleinhagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer. Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, 43(2–3):133–147, 2003.
- [7] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, and J. Fernández-Madrigal. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, AB, Canada, Aug. 2005.
- [8] E. Gat. On three-layer architectures. In D. Kortenkamp, R. P. Bonasso, and R. Murphy, editors, *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*. MIT/AAAI Press, 1997. RR n3552 26 Laugier, Fraichard, Garnier, Parmentchik and Scheuer.
- [9] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer. A multi-modal object attention system for a mobile robot. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1499–1504, Edmonton, Canada, 2005. IEEE.
- [10] S. Hüwel, B. Wrede, and G. Sagerer. Robust speech understanding for multi-modal human-robot communication. In *Proc. 15th Int. Symposium on Robot and Human Interactive Communication*. IEEE Press, 2006. to appear.
- [11] B. Krieg-Brückner, T. Röfer, H.-O. Carmesin, and R. Müller. A taxonomy of spatial knowledge for navigation and its application to the bremen autonomous wheelchair. In C. Freksa, C. Habel, and K. Wender, editors, *Spatial Cognition I – An interdisciplinary approach to representing and processing spatial knowledge*, pages 373–397. Springer, Berlin, Germany, 1998.
- [12] G.-J. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen. Clarification dialogues in human-augmented mapping. In *Proceedings of the ACM Conference on Human-Robot Interaction*, Salt Lake City, UT, USA, Mar. 2006.
- [13] B. Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119:191–233, 2000.
- [14] S. Lang, M. Kleinhagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proc. Int. Conf. on Multimodal Interfaces*, pages 28–35, Vancouver, Canada, 2003. ACM Press.
- [15] S. Li, B. Wrede, and G. Sagerer. A computational model of multi-modal grounding. In *Proc. SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL*, Sydney, Australia, July 2006. ACL Press.
- [16] S. Li, B. Wrede, and G. Sagerer. A dialog system for comparative user studies on robot verbal behavior. In *Proc. 15th Int. Symposium on Robot and Human Interactive Communication*, Hatfield, United Kingdom, September 2006. IEEE Press.
- [17] T. McNamara. Mental representations of spatial relations. *Cognitive Psychology*, 18:87–121, 1986.
- [18] O. Mozos, C. Stachniss, and W. Burgard. Supervised learning of places from range data using adaboost. In *Proceedings of the IEEE International Conference on Robotics and Automation*, Barcelona, Spain, 2005.
- [19] I. Posner, D. Schroeter, and P. Newman. Using scene similarity for place labeling. In *Proceedings of the International Symposium of Experimental Robotics*, Rio, Brazil, 2006. to appear.
- [20] T. Spexard, A. Haasch, J. Fritsch, and G. Sagerer. Human-like person tracking with an anthropomorphic robot. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1286–1292, Orlando, Florida, May 2006. IEEE.
- [21] E. Topp. *Initial Steps toward Human Augmented Mapping*. Licentiate thesis, KTH School of Computer Science and Communication (CSC), Royal Institute of Technology (KTH), Stockholm, Sweden. TRITA-CSC-A 2006:13, ISBN 91-7178-452-7, to be disputed 10/06/2006.
- [22] E. Topp and H. Christensen. Tracking for Following and Passing Persons. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, Edmonton, AB, Canada, Aug. 2005.
- [23] E. Topp and H. Christensen. Topological modelling for human augmented mapping. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Beijing, China, Oct. 2006.
- [24] E. Topp, H. Hüttenrauch, H. Christensen, and K. Severinsson Eklundh. Acquiring a shared environment representation. In *Proceedings of the ACM Conference on Human-Robot Interaction*, Salt Lake City, UT, USA, Mar. 2006.
- [25] D. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.
- [26] S. Wrede, J. Fritsch, C. Bauckhage, and G. Sagerer. An XML Based Framework for Cognitive Vision Architectures. In *Proc. Int. Conf. on Pattern Recognition*, number 1, pages 757–760, 2004.