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

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

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

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

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

General Terms
Human Factors, Design

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

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

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

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

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

2. RELATED RESEARCH

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

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

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

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

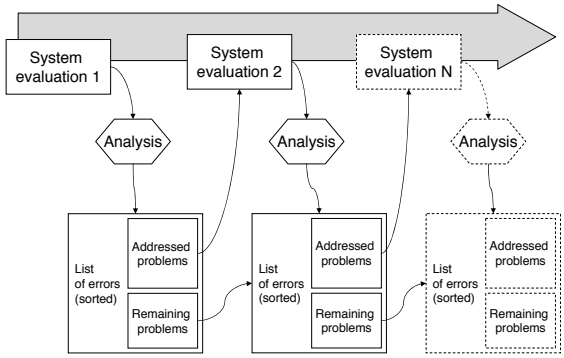

Figure 2: The role of miscommunication analysis in the design process.

3. MISCOMMUNICATION ANALYSIS

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

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

3.1 Purpose of the study

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

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

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

3.2 Dialogue model

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

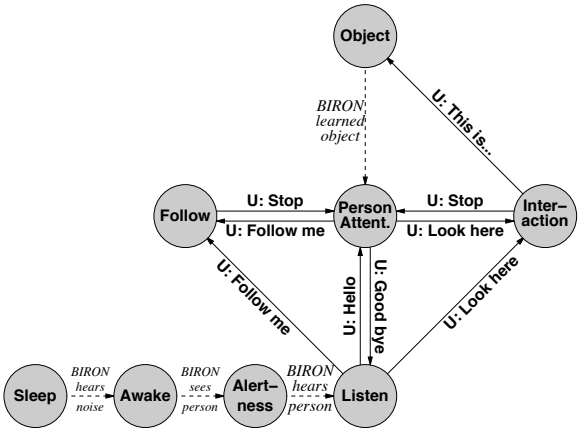

Figure 3: The dialog model described as a Finite State Machine.

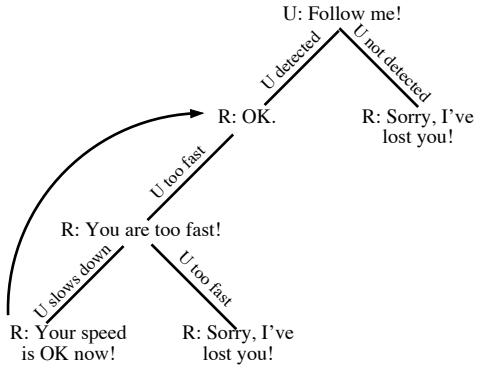

Figure 4: Sub-dialog of state "Follow".

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

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

3.3 Scenario and test procedure

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

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

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

3.3.1 Adaptation of dialogue patterns

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

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

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

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

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

3.3.2 Technical setup and test procedure

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

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

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

4. RESULTS OF THE ANALYSIS

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

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

4.1 Users' system knowledge

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

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

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

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

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

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

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

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

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

4.2 Feedback problems

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

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

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

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

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

4.3 Irrelevant response

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

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

4.4 Reference

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

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

5. DESIGN IMPLICATIONS

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

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

5.1 Information given to the user

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

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

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

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

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

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

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

5.2 User priming

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

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

5.3 Monitoring communicational success

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

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

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

5.4 Recovery strategies

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

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

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

5.5 Discussion

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

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

6. CONCLUSIONS

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

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

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

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

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

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

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

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

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

7. ACKNOWLEDGMENTS

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

8. REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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