John Benjamins Publishing Company

This is a contribution from *Interaction Studies 10:3*

Improving HRI design by applying Systemic Interaction Analysis (SInA)

Manja Lohse, Marc Hanheide, Karola Pitsch,
Katharina J. Rohlfing & Gerhard Sagerer
University of Bielefeld, Germany

Social robots are designed to interact with humans. That is why they need interaction models that take social behaviors into account. These usually influence many of a robot's abilities simultaneously. Hence, when designing robots that users will want to interact with, all components need to be tested in the system context, with real users and real tasks in real interactions. This requires methods that link the analysis of the robot's internal computations within and between components (system level) with the interplay between robot and user (interaction level). This article presents Systemic Interaction Analysis (SInA) as an integrated method to (a) derive prototypical courses of interaction based on system and interaction level, (b) identify deviations from these, (c) infer the causes of deviations by analyzing the system's operational sequences, and (d) improve the robot iteratively by adjusting models and implementations.

Keywords: analysis tools, user studies, autonomous robots

1. Introduction

The field of human–robot interaction (HRI) brings together researchers from various disciplines such as human–computer interaction, computer science, psychology, and cognitive science. While each discipline contributes its own methods, many researchers focus on specific challenges such as navigation and perception of humans, and they evaluate their individual components in terms of success rates and performance metrics. Despite the value of such findings, models of components also have to be evaluated in the context of the system as a whole, because the complexity of processes in a system, the perceptual limits of the system sensors, and, in particular, the interaction and task context have to be taken into account. A component that works well when tested independently may turn out to be problematic once it has been integrated into a system with other components based on different models. Moreover, whereas

Interaction Studies 10:3 (2009), **298–323**. DOI 10.1075/is.10.3.03loh
ISSN 1572–0373 / E-ISSN 1572–0381 © John Benjamins Publishing Company

the underlying model of the component might appear adequate from a technical point of view, it may well be inappropriate for interactions with a user under specific conditions. Therefore, the field still lacks a systemic analysis approach that simultaneously addresses the system and the interaction level of HRI. The system level includes the components and their interplay, whereas the interaction level describes the interplay between the system and its user in the context of a specific task.

In this article, we introduce Systemic Interaction Analysis (SInA) as our approach to analyzing system and interaction levels in an integrated manner. With the help of SInA, we attempt to improve HRI by understanding what humans do when interacting with an autonomous robot, what might motivate their behavior, what the robot does, and what happens within the system. Our aim is not to conduct in-depth analyses of the user's motivations and behaviors from a psychological point of view, but to provide a careful description of the interaction that will enable us to determine relations between the user's and the system's behavior.

To understand the value of SInA, one has to consider the multidisciplinary development process of interactive robot systems. Usually, this development is guided by insights from human–human interaction that are entered into models for the implementation of certain interactive behaviors in autonomous robots. Although developers do their best to make these implementations as advanced and robust as possible, their straightforward and direct realization often leads to unforeseen effects. We could term this kind of implementation strategy an *open-loop* design. However, if interaction with a robot is to be successful and acceptable, these implementations need to be assessed and revised in a *closed-loop* fashion by carrying out thorough interaction studies. SInA is our attempt to face the challenge of closing the loop in a systematic way by explicitly relating interaction phenomena observed *externally* to the *internal* operation flow of the system. SInA is based on combining task analysis (TA) and ethnomethodological conversation analysis (CA) to form an interaction analysis framework. In the following related work section we discuss the theoretical basis. Thereafter, we describe SInA in depth (see section "Systemic Interaction Analysis (SInA)"). SInA can be applied to a range of scenarios and robots. However, HRI according to Steinfeld et al., (2006) is task-driven. That is why our analysis focuses strongly on the scenario and the robot system of the user study (see section "Scenario, user study, and robot system"). In this article, we describe the structure of SInA and apply it to a case study of a specific task, namely location-learning (see section "SInA case study"). The findings from this case study are then compared with findings from an analysis of the follow behavior of our robot conducted with the same method (Lohse, Hanheide, Rohlfing, & Sagerer, 2009).

These case studies illustrate how SInA is applied and the benefits of its application to iterative robot design.

2. Related work

This section introduces TA and ethnomethodological CA, describes how they relate to HRI, and shows how they are combined in our interaction analysis approach.

TA has been used in market research, product testing, and also in human–computer interaction for a long time. In recent years, it has also been introduced to HRI. For example, Adams (2005) applied it in the context of situational awareness and user roles. Hüttenrauch, Green, Oestreicher, Norman and Severinson Eklundh (2004) have proposed TA as a means to identify user's work procedures and tasks when interacting with a mobile office robot, the physical design requirements, function allocation and the relation of the work between user and robot, and user's expectations. The authors approached these goals through interviews and focus groups. In a similar vein, Kim and Kwon (2004) have applied TA for system design in HRI. However, none of these authors has focused on TA in the evaluation process, and no data-driven approaches based on user studies have been proposed so far. We believe that TA is also useful in these respects. In particular, hierarchical TA as developed by Annett and colleagues (see Stanton, 2006) is a useful tool for HRI because it can be used to define tasks and subtasks and their interrelationships. The method links hierarchically structured subgoals by plans determining when a subgoal is triggered. These subgoals have to be met in order to attain the goals. One main advantage of the hierarchical TA structure is that it allows further subtasks to be added at any time, making it applicable to all scenarios and robot systems.

Another key method in SInA is ethnomethodological conversation analysis. CA has developed a methodology for the fine-grained empirical-qualitative analysis of audio and video data in order to study the sequential organization of human interaction. More recently, a central focus has been multimodal aspects of communication (Goodwin, 2000; Sacks, 1992). Lately, a small group of researchers conducting CA have begun to consider HRI. The few available findings support the use of CA on two levels: (a) to study human interaction in authentic situations and to generate a model for designing the communicational interface of the robot that uses statistical methods to evaluate the interaction with the human user (Kuzuoka et al., 2008; Yamazaki et al., 2008); and (b) to study the interaction between human and robot in experimental settings (see Muhl, Nagai, & Sagerer, 2007 for a sociological approach).

Our approach attempts to combine both levels: to study HRI in order to reveal fine-grained details of the unfolding sequential organization in which the robot's actions are embedded while simultaneously pointing out interaction-related problems that might exceed the behavior of the robot implemented so far and suggest new ways of advancing the robot's model of behavior.

In its traditional form, CA serves to analyze the processes and the structure of an interaction. The processes fill out the structure by describing procedures and methods of the interaction partners. While CA is an appropriate instrument for our description of processes, we use TA to analyze the structure of the interaction. In our research, the structure is provided by the tasks users have to complete with the robot. Much knowledge about them is derived from the technical implementation. Therefore, while we suppose that both methods would lead to the same result, it is much more efficient to identify the structure with TA. This is because TA includes the given knowledge, whereas CA would need to analyze the interaction from scratch by conducting a full video analysis. Moreover, in contrast to CA, TA permits a quantification of the results.

Both methods, CA and TA, are combined to form an interaction analysis process. The resulting SInA approach shares many characteristics with the traditional interaction analysis applied in human–human interaction, because both are inspired by CA approaches, and both enrich classic CA with ethnographic methods. In addition, the interaction analysis approach as described by Jordan and Henderson (1995) is interdisciplinary, and it is applied to the empirical investigation of interaction between interlocutors with each other and their environment. We also share the predominant research interest of interaction analysis in investigating how people make sense of each other's actions, and how this can be seen in their actions (Jordan & Henderson, 1995). The main difference between our research and traditional interaction analysis is that an autonomous robot is participating in the interactions analyzed here.

Burghart, Holzapfel, Haeussling, and Breuer (2007) have also introduced an interaction analysis approach to HRI. Their "Interaction Analysis Protocol" (IAP) is a tool to analyze the metrics of the interaction in HRI based on video and transcriptions. The IAP consists of six layers of information such as transcripts of the verbal and nonverbal actions of both participants, phases in the interaction, and problems and notable incidents. The authors analyze problems in the interaction in order to derive advice on how to solve them. In this respect, the goal of the approach seems to be closely related to our research. However, the IAP tool does not take the system level into account. Therefore, we assume that it cannot be used to trace problems back to certain system components as envisioned in closed-loop design. Moreover, the level of analysis seems to differ between the approaches. Whereas Burghart and colleagues break an interaction down into "phases" (opening, negotiation, and ending of the interaction), our units of analysis are tasks identified with the help of TA. The advantages of this will be explained below.

3. Systemic Interaction Analysis (SInA)

The goal of SInA is to close the loop between technical implementation and user studies by directly incorporating evaluation findings that are based on the analysis of system and interaction level into the system. To approach this goal, we conduct user studies in which inexperienced users interact with an autonomous robot. The evaluation of our autonomous system focuses mainly on understanding and assessing (a) how it performs in collaboration with humans and (b) why it performs in the observed way. This process requires answers to the following concrete questions:

- What do the users do?
- What happens within the system?
- What does the robot do?

To answer these questions, we videotape the interactions and record the system's operation flow by exploiting centralized logging functions in the respective components to produce time-synchronous log files. We then analyze the data acquired in the studies following the SInA procedure (see Figure 1).

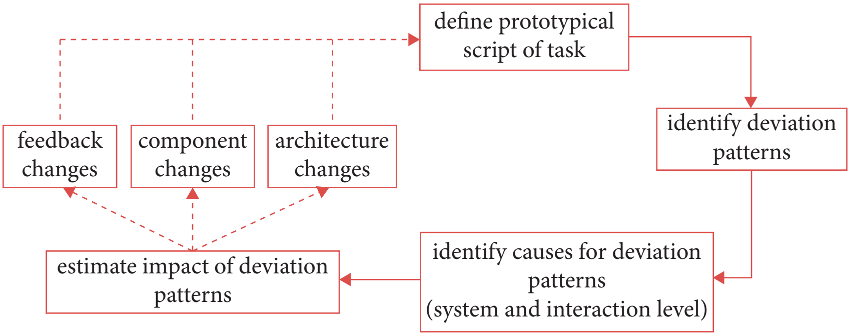

define prototypical script of task
identify deviation patterns
identify causes for deviation patterns (system and interaction level)
estimate impact of deviation patterns
feedback changes
component changes
architecture changes

Figure 1. SInA cycle

3.1 Prototypical interaction script

Before starting the actual SInA analysis, we apply hierarchical task analysis to identify all tasks and subtasks (see the following section for all tasks in our scenario). We then select a task and take a careful look at all its instances to define the *prototypical interaction script*. The script is identified on the basis of the envisioned interaction model of the developer and its application and restrictions in real-world situations observed in video data from user studies. As in traditional interaction analysis, the script is developed in video sessions with interdisciplinary participation (Jordan & Henderson, 1995). TA contributes to the process by providing a description of the task. The description is further specified with the help of ethnomethodological CA, which provides insights into the communicative surface of the interaction, the sequential organization of the events, and the precise coordination of the different communicational resources.

3.2 Deviation patterns

In the second step of SInA, we identify cases in which the interaction deviates from the prototypical script. Deviations are to be expected if a component is tested in an integrated system with real users and its model needs to be adapted. Moreover, deviations may be caused by perceptual restrictions of the system that become apparent in user studies. Components also have to be adapted to these.

Deviating cases are observed on the interaction level, and their causes are traced back to the system level on which the components involved are identified. This constitutes the core idea of SInA. In order to verify that phenomena have not occurred by coincidence, we search for further examples of each phenomenon. Deviations that occur only once are not included in the SInA procedure. However, they are noted for latter analysis within a CA approach.

In the next step, we define groups of deviating cases that we call *deviation patterns*. Each deviation pattern includes cases that are similar in terms of what the users do, what happens within the robot, and what the robot does. These three factors indicate possible causes of deviation patterns. We derive a categorization of the patterns by clustering them according to the robot's functions.

Within this second step, we obtain quantitative measures of the occurrence frequencies of the patterns with the help of TA. These provide an estimation of the relevance of any given deviation. This relevance is also determined by a deviation pattern's influence on the further course of the interaction, and this is analyzed as well. Influence is high if the deviation pattern interrupts the interaction completely or for a long time; it is low if the problem can be resolved quickly or the user does not even notice the deviation. Moreover, a comparative analysis of all tasks provides information on the impact of a phenomenon. If a deviation occurs in many tasks, its relevance is higher than if it occurs in one task alone.

3.3 Learning from deviation patterns

In the third step, the knowledge about the patterns and the underlying system design problems is used to address the deviations in the further development process. This results in a need to either (a) redesign system components (what happens within the system), (b) influence users' behavior by designing appropriate feedback, or (c) consider a redesign of the system architecture. Although these changes may be rather short-term (next iteration), it may also be necessary to include long-term improvements of interaction models (see Figure 2).

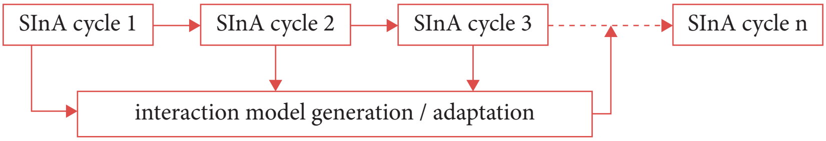


Figure 2. Short-term and long-term effects of SInA

Finally, the changes have to be evaluated. This step can be achieved only by reiterating the whole SInA procedure. Therefore, as Figure 1 shows, the approach is based on a cyclic, iterative model. The prototypical interaction script, which might include technical restrictions, has to be reviewed in each iteration.

This section has introduced SInA from a theoretical point of view. In the following, we present a case study, the scenario, and the robot system on which it is based.

4. Scenario, user study, and robot system

In a case study, we applied the SInA method to the home tour scenario. The home tour focuses on multi-modal HRI to enable the robot to learn about the domestic environment and its artifacts, the appearance and location of objects, and their spatiotemporal relations. With these abilities, the robot can adapt to new settings like a user's home. The environments have to be explored jointly with the user. The home tour focuses on the tutoring of the robot. It includes five major tasks (see Figure 3).

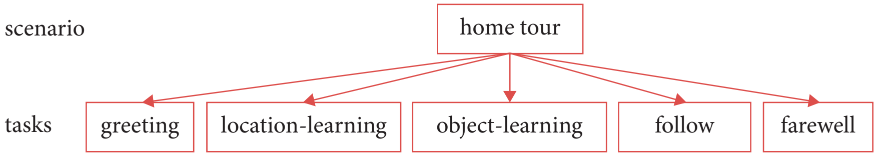


Figure 3. Home tour tasks

The robot's capabilities for natural interaction in the home tour comprise: understanding of spoken utterances, co-verbal deictic reference, verbal output, referential feedback, and person attention. Furthermore, the robot has to be able to follow the user.

To evaluate these capabilities, we conducted a user study consisting of two iterations with 24 subjects in August and November 2007. All participants were German native speakers who interacted with the robot BIRON (see below) in German. In the first study in August 2007, some of the participants were university students (average age 25.6 years; 7 female, 3 male); in the second study in November 2007, they were mainly older adults (average age 45.5 years; 5 female, 9 male). Even though participants received a small reward for participation, their main motivation was to get to know the new technology. Therefore, they might have been more interested in technology than average people. This was also displayed by their knowledge of computers (average 3.2 on a scale ranging from 1 [no knowledge] to 5 [expert knowledge] in both studies). Despite their strong interest, participants were, nonetheless, inexperienced in interacting with robots (average 1.4 on a scale ranging from 1 [no experience] to 5 [a lot of experience]). This was taken into account when designing the following study procedure. First, participants were welcomed and introduced to the study. They answered a questionnaire tapping demographic data and their experience of interacting with robots. Afterwards, they were trained to use the speech recognition system, that is, they were instructed about the proper placement of the headset microphone and were asked to speak some phrases to familiarize themselves with its use. The recognition results were displayed in verbatim on a laptop computer. Afterwards, participants were taken to the room where the robot was waiting ready for operation. They were assisted during the first contact in order to reduce hesitant behaviors. They were handed a training script to practise interaction with the robot. During this initial training session, the experimenter also instructed the users on how to pull the robot when it got stuck. After the training session, participants returned the training script before performing the main task, namely, to teach the robot objects and locations in several rooms. The instruction for this main task was:

- guide the robot through the apartment, that is, from the living room to the dining room via the hall (see Figure 4)
- show and label the living room and the dining room
- show green armchair in the living room and the table in the dining room (in the August session); show the cupboard in the living room and the floor lamp in the dining room (in the November session)

During this part of the interaction, the experimenter intervened only when asked by participants. The whole interaction including the training session was videotaped. Afterwards, participants were interviewed about the interaction in general and the robot's visual display in particular. They then completed a second questionnaire containing items on their liking of the robot, attributions made to it, and its usability.

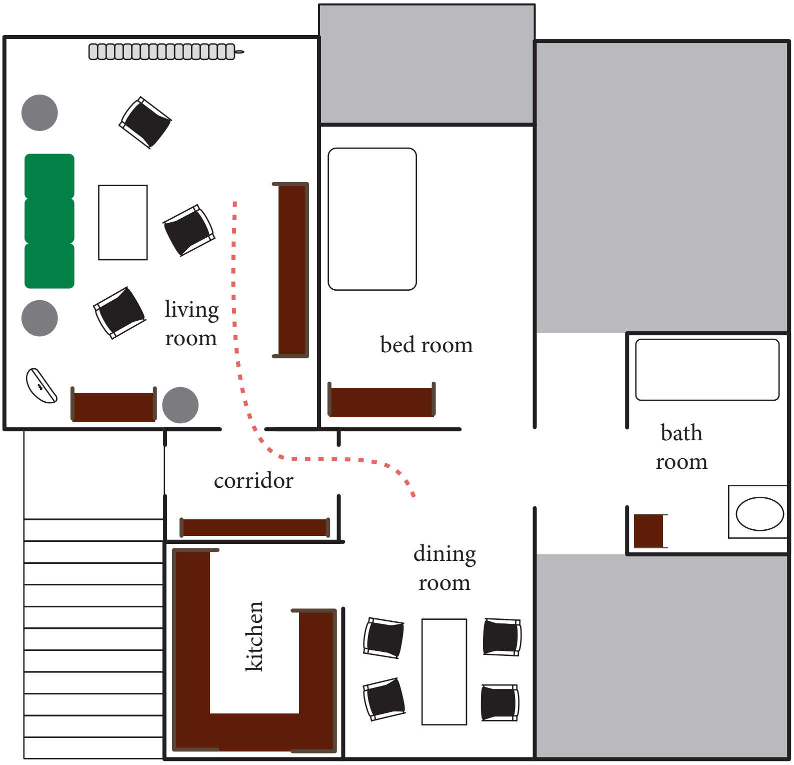


Figure 4. Apartment (red line: path the robot had to be guided)

The robot used for the trials is called BIRON (BIelefeld Robot companiON, see Figure 5). BIRON is based on a Pioneer PeopleBot platform. A Sony EVI D 31 pan-tilt color camera mounted on top of the robot at a height of 142 cm is used to acquire images of the upper body part of humans interacting with the robot and to focus on referenced objects. An additional camera is used to capture hand movements to recognize deictic gestures. A pair of AKG far-field microphones is located below the touch screen display at a height of approximately 107 cm. These enable BIRON to localize speakers. Finally, a SICK laser range finder is mounted on the front at a height of 30 cm. It measures distances within the environment in order to detect pairs of legs and to avoid obstacles while driving.

Many of the components serve for the robust perception of the interaction partner, which is an essential prerequisite for the interaction. The actively controlled pan-tilt camera to detect faces, the stereo microphones for voice direction perception, and the laser scanner to detect pairs of legs enable the robot to perceive and continuously track multiple people in real-time.

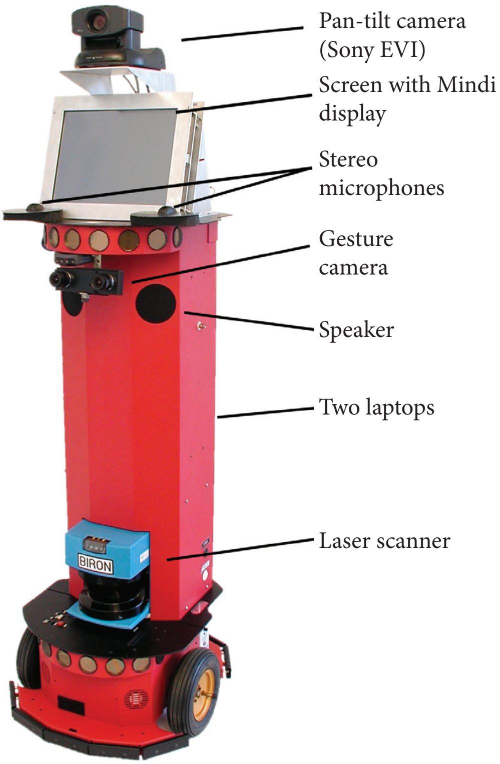


Figure 5. The robot BIRON

To use its different abilities in a way that is explicable and clear to users, the robot employs behavioral states, namely "Following," "Location Learning," "Object Learning," "Person Attention," and "Alertness." More details about the architecture and design principles can be found in Hanheide and Sagerer (2008).

The robot also features a grounding-based dialog component that manages the interaction with the user (Li, Wrede, & Sagerer, 2006). It employs not only verbal in- and output but also visual feedback using a virtual character called Mindi that resembles BIRON. Mindi is displayed on the robot's screen. The Mindi pictures relate directly to the robot's state, for example, if the robot is following, an animation is shown in which the Mindi has a happy face and is also walking; if the robot is learning a location, Figure 6 is displayed.

To record all relevant information on the system level, we exploited the robot's event-driven architecture (Hanheide & Sagerer, 2008) and standardized logging tools.[1] These enable us to generate time-synchronous logging of the internal processing of the different components. The system logs are translated

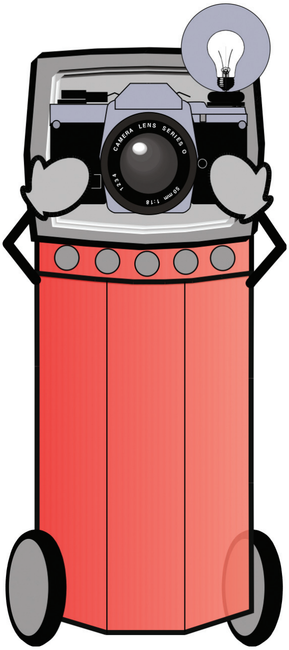


Figure 6. Screen character Mindi during location-learning task

automatically into ELAN[2] (Linguistic Annotator) files and synchronized to video recordings and manual annotations. With the help of ELAN, multilayer annotations are created and hierarchically structured. Table 1 provides an overview of all logs and annotations.

Table 1. Manual annotations and automatic logs linking system and interaction level

Manual Annotations	**Automatic Logs**
– Speech: participant to robot	Actions
– Speech: participant to experimenter	– Arbitration commands (e.g., hardware access)
– Speech: experimenter to participant	– Speech output of robot
– Gestures of participants	– Mindi (screen)
– Rooms shown	– Motor commands
– Objects shown	Perception
	– Perceived speech input of participants
	– Perceived gestures of participants
	– Location recognition
	– Tracked persons
	– Obstacle detection
	States
	– Dialog (grounding)
	– General behavioral state

For our analyses, we synchronized and incorporated the data into one ELAN file for every participant, so that we could work with all information at once (see Figure 9). In principle, the method can be applied using other annotation tools as well.

5. SInA case study

As described in Section 4, one task within the home tour scenario is to teach the robot locations. Here, we choose this task to illustrate how the Systemic Interaction Analysis is applied. At the end of the section, we compare the results with analyses of the robot's follow behavior (Lohse, Hanheide, Rohlfing, & Sagerer, 2009) and work out implications for robot design.

5.1 Prototypical interaction script

First of all, we defined a *prototypical interaction script* for the location-learning task based on a data driven hierarchical task analysis. On the interaction level, we studied the questions "What do users do" and "What does the robot do?" The answers to these questions are depicted in Figure 7.

task	location-learning		
	user		robot
subtasks	indicate that a room is shown and label the room	>	recognize that a room is shown and the name of the room ∨ acknowledge both of the above verbally and with Mindi ∨ 360° turn to refine the map of the room

Figure 7. Task description of location-learning

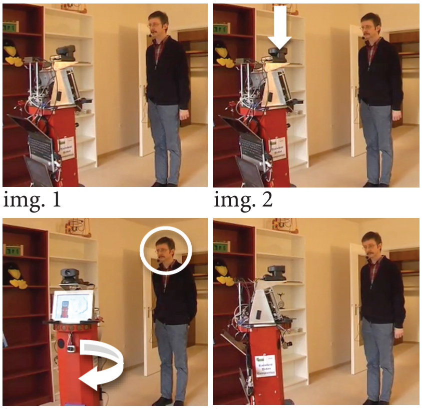
If the robot recognizes the location-learning command, it acknowledges it verbally. At the same time, Figure 6 is displayed on its screen. In order to acquire a decent representation of the room, the robot conducts a 360° turn. This enables it to improve its metric mapping using SLAM (Guivant & Nebot, 2001). Furthermore, it uses its laser scanner to register the coarse layout of the room such as its size and its geometrical shape (Christensen & Topp, 2006). The researchers can observe the robot and human behavior discussed so far in the video recordings. In the following, we describe internal processes based on the system level of the robot.

On the system level, several prerequisites have to be fulfilled to conduct a location-learning task:

- The system is in a state to accept the location-learning command
- The overall system state is consistent (all components can accept the command)
- The robot perceives the person stably
- The system understands the user's utterance and interprets the command correctly

If these prerequisites are fulfilled, the robot state switches to location-learning. A prototypical interaction sequence for the location-learning task is presented in Fragment 1 (Figure 8). In this example, the user had maneuvered BIRON to the center of the living room, made the robot stop, and—positioned face-to-face—uttered "↑Biron; (.) THIS is the living room" (img. 1). BIRON lowered its upper camera (img. 2), then began to turn slowly while repeating the new information "this is the living room then."

Not only was the task carried out as anticipated, but the user could also be seen to acknowledge the robot's behavior as appropriate: following the robot's utterance and the beginning of its turning motion, he nodded while assuming an acknowledging facial expression (img. 3). He waited until the robot had finished inspecting the room and returned to its original position before performing any further activities (img. 4). Note that the precise timing of the robot's camera movement and body turning in relation to its verbal utterance were achieved by successively fine-tuning the system to achieve a sequence of actions in which the robot could be seen to perform a logical suite of actions: inspecting the room before commenting on it.


img. 1
img. 2
img. 3
img. 4

01 U: ↑Biron; (0.1) DAS ist das WOHNzimmer
↑Biron; (.) THIS is the LIVING room
*img.1
02 P : (1.8)
03 B : |(0.2) d|as ist also das wohnzimmer;
this is then the living room
|cam ↓ |
|turns… 20°
*img.2
04 U: |nods acknowledging |waits
B : … …150°
*img.3 *img.4

Figure 8. Fragment 1 (Nov. VP09, 14:00 min)[3]

As described before, we do not assume that a prototypical situation is "perfect" or comparable to an ideal situation in human–human interaction. Rather, we take restrictions of the system into account. In the case analyzed here, one such restriction is that the robot might lose the interaction partner while turning. Due to the hardware design, the robot is "blind" behind its back. Hence, it can no longer track or detect the interaction partner. Instead of concealing this implementation drawback, we decided to tackle it by giving appropriate feedback. Though the robot will usually not be able to track a person successfully, the idea is that feedback lets the user know how to re-initiate the interaction. In this case the robot asks the user to say "hello" if she wants to continue the interaction.

5.2 Analysis of deviation patterns

In the next step, we observe cases on the interaction level in which the interaction deviates from the prototypical script and trace back their causes to the system level. We sum these cases up to form *deviation patterns*.

Altogether, 128 location-learning sequences were analyzed. 36 (28%) of these were categorized as being in accordance with the prototypical interaction script. Prototypical interactions include the ones in which the robot had not lost the user while turning (11), as well as ones in which the robot lost the user and asked him or her to say "hello" if he or she wanted to continue the interaction (25). The remaining 92 cases were deviating cases. With the help of the method described above, we succeeded in categorizing these to five deviation patterns that explained 86 (93%) of the non-prototypical sequences. The deviation patterns can further be categorized into three groups of robot functions: speech recognition, person perception, and state-related patterns. They are described in the following. Table 2 analyzes them according to what the users do, what happens within the system, what the robot does, the number of occurrences, and their influence on the interaction.

6. Speech understanding

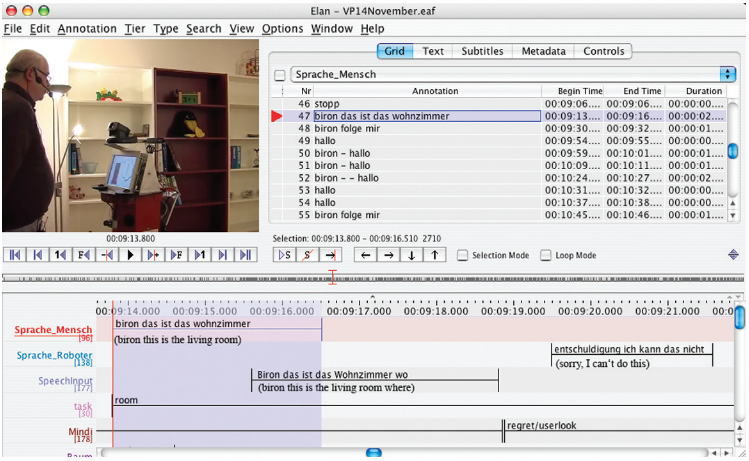
This category contains only one deviation pattern in relation to learning locations, namely, errors in speech recognition.

Errors in speech recognition:
Current speech recognition technology is still unable to achieve 100% accuracy, particularly in the speaker-independent recognition needed in the home tour scenario. Therefore, errors in speech recognition are a major cause of deviations from the prototypical interaction script. Although the robot features a speech

Table 2. Deviation patterns for location-learning task

Pattern	User	System internal	Robot	#	Influence on interaction
Speech understanding					
Errors in speech recognition	Utters a location-learning command	(a) Input cannot be interpreted at all	(a) Asks user to repeat	5	(a) Users repeat command
		(b) Input is interpreted in a wrong way	(b) Starts a wrong action	31	(b) Users try to resume task
Person perception					
Third person	Utters a location-learning command	Mistakenly classified a third-person situation	Does not react to user	5	Users wait; try to resume task
Person lost	Utters a location-learning command	No person perceived according to person model	(a) Does not react	1	(a) Users wait for a reaction; if nothing happens, try to attract robot's attention
			(b) Asks person to register	4	(b) Users say "hello" again; interaction continues
States					
Unfinished or wrong state	Asks robot to learn a room before completing some other action	Room cannot be learned; utterance is interpreted within the current system state	Asks user if she wants to do something else, tells her she has to say stop before she can do so	15	Users wonder why robot cannot learn location; say "stop" to finish former action
Action incomplete	Teaches a room, does not wait for robot to turn before starting a new action	Turning is interrupted	Starts new action	25	Robot cannot refine map, users do not notice deviation

understanding component that can initiate clarification turns ("Pardon me?") in case of detected errors in speech recognition, it might also simply interpret the utterance incorrectly and, as a consequence, trigger an unexpected behavior, for example, when the robot asked the user if she wanted to know where they were instead of learning the location. This happened most frequently because the speech recognition perceived the sentence "BIRON, das ist das Wohnzimmer WO" instead of "BIRON, das ist das Wohnzimmer" ("BIRON, this is the living room WHERE" instead of "BIRON, this is the living room"). The interrogative in the sentence led the robot to believe that the user was asking for a label of the current location. One example of this situation is depicted in Figure 9. It was noticeable that speaker dependence of speech recognition was very high in the user study. Only 2 out of 24 subjects triggered 16 of the 36 deviation sequences (44%).


Elan – VP14November.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Metadata Controls
☐ Sprache_Mensch
Selection Mode
Loop Mode

Nr	Annotation	Begin Time	End Time	Duration
46	stopp	00:09:06....	00:09:06....	00:00:00....
47	biron das ist das wohnzimmer	00:09:13....	00:09:16....	00:00:02....
48	biron folge mir	00:09:30....	00:09:32....	00:00:01....
49	hallo	00:09:54....	00:09:55....	00:00:00....
50	biron – hallo	00:09:59....	00:10:01....	00:00:01....
51	biron – hallo	00:10:09....	00:10:11....	00:00:01....
52	biron – – hallo	00:10:24....	00:10:27....	00:00:02....
53	hallo	00:10:31....	00:10:32....	00:00:00....
54	hallo	00:10:37....	00:10:38....	00:00:00....
55	biron folge mir	00:10:45....	00:10:46....	00:00:01....

00:09:13.800

Selection: 00:09:13.800 – 00:09:16.510 2710

00:09:14.000 00:09:15.000 00:09:16.000 00:09:17.000 00:09:18.000 00:09:19.000 00:09:20.000 00:09:21.000

Sprache_Mensch [96]: biron das ist das wohnzimmer (biron this is the living room)

Sprache_Roboter [138]: entschuldigung ich kann das nicht (sorry, I can't do this)

SpeechInput [177]: Biron das ist das Wohnzimmer wo (biron this is the living room where)

task [30]: room

Mindi [178]: regret/userlook

Raum

Figure 9. ELAN annotation of situation with an error in speech recognition

7. Person perception

Person perception includes two deviation patterns, namely person lost and third person perceived.

Person lost:
A function that is particularly relevant for a successful interaction is the robust tracking of the user. If the person is "lost" or not yet registered in the system, the robot will not react when talked to or it will ask the user to re-register.

Third person:
Considering the fact that in typical domestic environments the robot is not the only interlocutor for a human, the robot must be aware of other potential humans in order to discern whether it is being addressed or not. Therefore, the robot checks the face orientation of surrounding humans and only accepts verbal input from

people facing it directly in line with the hypothesis that humans usually address others by looking at them. This is why the robot did not react to commands if it detected a so-called "third person." If a third person was perceived by mistake, this led to a deviation pattern because the robot, viewed from the interaction level, did not react to the user and did not provide an explanation for its behavior.

8. Robot states

The third group of deviation patterns concerns robot states. For the location-learning task, this includes unfinished or wrong states and incomplete actions.

Unfinished or wrong states:
As described above, the robot operates in various states. All states have to be finished before starting a new action. For example, the robot has to finish the follow state before it can learn a location. If a state was not finished, the grounding based dialog did not accept the new action.

Incomplete actions:
In contrast to "unfinished or wrong states" that focus on actions before the location-learning sequence, this category describes the incomplete task itself. Location-learning tasks were usually not completed if the robot did not turn. This happened when the robot could not turn because the carpet impeded it or because the users moved on immediately to the next task, not knowing that turning allowed the robot to obtain a better representation of the environment.

8.1 Comparison to the follow task

One major advantage of SInA is that it can be used to compare new findings with analyses of other tasks. This tells us whether deviation patterns occur in more than one task, and if they have already been addressed in other tasks. Furthermore, we receive a more reliable estimate of how severe a deviation pattern is, and how much it influences the interaction. This article focuses on the comparison of the follow task and the learning-location task. These findings are depicted in Figure 10.

The follow task was analyzed according to the same procedure as presented here (see Lohse, Hanheide, Rohlfing, & Sagerer, 2009). As a result, we also obtained a task description and deviation patterns. Altogether, 264 follow sequences were analyzed, 219 being deviating cases, of which 98% (215) were explained by 10 deviation patterns.

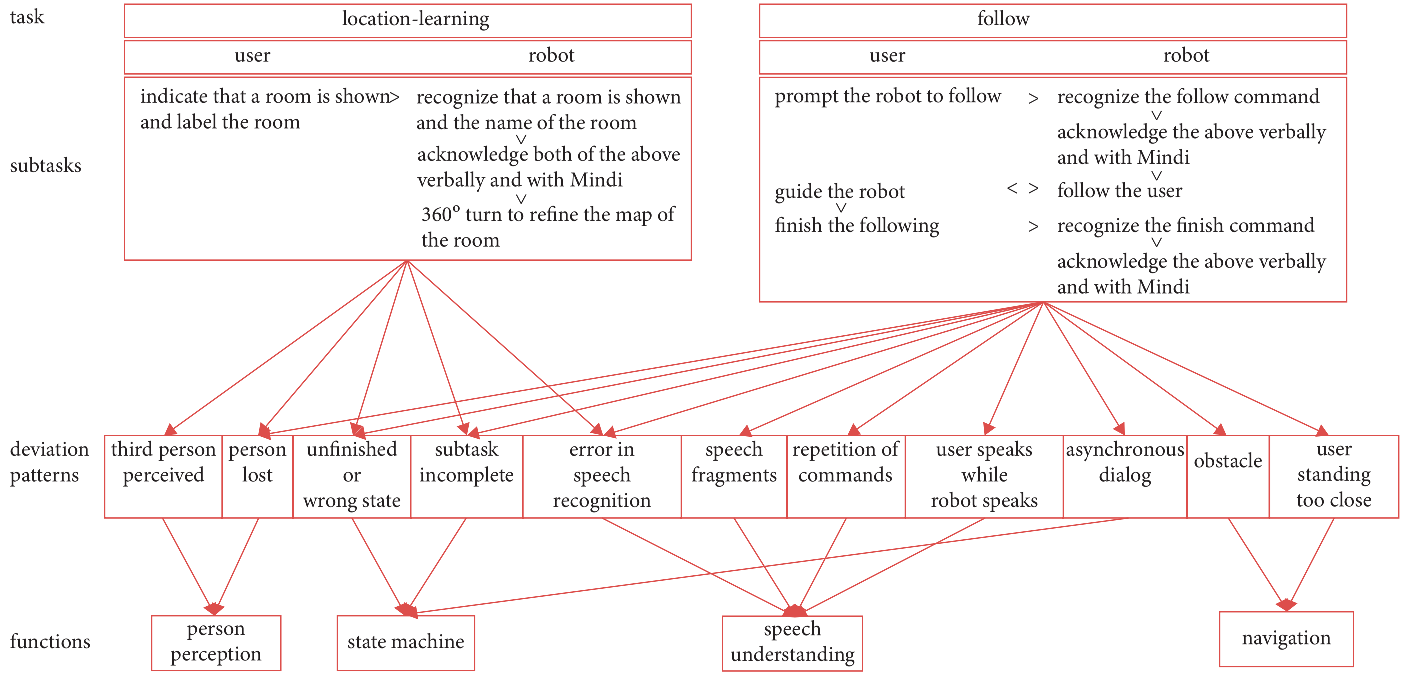


Figure 10. Task analyses for location-learning and following with resulting deviation patterns categorized by the robot's functions

9. Speech understanding

As in the location-learning task, speech recognition was a major challenge in the follow task. However, we found that speech-recognition-related deviation patterns were more diverse here. While identifying only one deviation pattern in the location-learning task, we differentiated four patterns for the follow task. Whereas the pattern "error in speech recognition" was equal for both tasks, in the follow task, "speech fragments," "repetition of commands", and "person speaks while robot speaks" also occurred. Speech fragments were noises the robot interpreted as utterances. These occurred in the follow task, because the robot itself and the human made a lot of noise while walking. In the location-learning task, in comparison, both stood still.

Repetition of commands also occurred only in the follow task. This included cases in which the user gave a follow or a stop command twice with only a short break. Hence, following seems to be much more time-critical than learning locations. Whereas the users gave the robot time to look at a room and to learn its name, they expected it to follow and especially to stop immediately.

Another deviation pattern that occurred only in the follow task was "person speaks while robot speaks." The inability of the robot to listen while it speaks causes problems, especially with short commands like "follow me" and "stop." As soon as the robot says, for example, "Pardon me," it might not hear the command. The situation was different in the location-learning task, because it used longer robot utterances. Therefore, the robot did not miss commands completely.

10. Person perception

Comparing person perception for both tasks (50 occurrences [23%] follow task vs. 5 occurrences [6%] location-learning task) indicates that perception worked well as long as the person did not move and lighting conditions were stable. Therefore, person perception should be adapted to the task the robot is working on.

11. Robot states

In both tasks, the state-oriented architecture of the robot was responsible for deviation patterns. The analysis showed that roughly the same percentage of "unfinished or wrong state" cases occurred. The causes and impact of this deviation pattern were the same in both cases: an action that was not finished before the tasks had to be completed with a "stop" command. In contrast, the impact of the next pattern—namely, action incomplete—was different in the two cases. In the learning-location task, the action was incomplete if the robot had not turned and refined its map of the room. The user might not even have noticed that the action was incomplete. In the case of the follow task, the impact on the course of the interaction was much stronger. The next task could not be started before the follow task was explicitly finished with a "stop" command. Since this was not consistent for all tasks, it sometimes confused the users. Another deviation pattern that occurred only in the follow task was asynchronous dialog. This pattern describes cases in which the interaction partner was lost but the dialog system was not yet aware of this. The causes of this pattern are to be found within the design of the follow task. In the system architecture, the follow behavior was not yet designed according to the state sequence protocol (Peltason et al., 2009). This particular architectural improvement had already been made to the location-learning task. Consequently, the pattern of asynchronous dialog did not occur in location-learning, confirming the appropriateness of this modification.

12. Navigation

Navigation includes deviations that occurred while the robot was being guided. Hence, these patterns cannot be compared to the location-learning task. Nonetheless, we shall describe them briefly. First, the robot might be blocked by an obstacle. The second deviation pattern in this category is "person standing too close." As the robot maintained a certain security distance to the user, it stopped driving as soon as it came too close to her. If she did not interpret this correctly and did not increase her distance to the robot but instead waited for the robot to approach, the interaction got stuck.

12.1 Implications for robot design

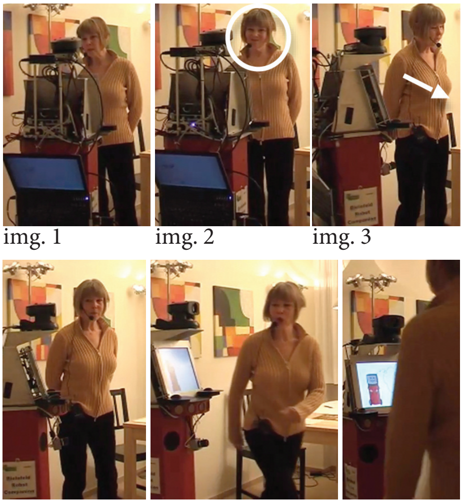
After having identified and compared the deviation patterns, the final step in the cycle is to tackle them in order to improve the interaction. Some examples shall be discussed here. The most important issues regarding the number of occurrences when analyzing the location-learning task were errors in speech recognition, unfinished or wrong states, and incomplete actions (see Table 2). These were equally important in the follow task.

As pointed out before, errors in speech recognition were crucial for both tasks. As we found strong speaker dependence, speaker adaptation and task-specific language models can probably achieve better speech recognition results.

Unfinished and wrong states can be avoided only by changing the design of the tasks that cause these problems. Most of the incidents were caused by incomplete follow actions. As the comparison has shown, it is necessary to change the design of this task and adapt it to the other tasks.

Whereas the TA discussed so far has been able to identify deviation patterns and their causes, methods of ethnographical CA are able to provide insights into details of the concrete interaction behavior (Goodwin, 2000; Sacks 1992). We shall demonstrate this with respect to the robot's 360° turn to build up a representation of a room.

Similar to the first case in Figure 8, in the situation depicted in Figure 11, the user had maneuvered BIRON to stand face-to-face with her in the center of the (dining) room and suggested, "THIS is the dining room; ↑BIron;" (01, img. 1). BIRON lowered its camera a little, uttered "nice" (02), which elicited a smile from the user (img. 2). A fragment of a second later, BIRON proceeded to repeat the new information "this is your dining room then" while turning slowly (02–03), and the user ratified "yes" and turned away (img. 3).



```
01 U: DAS (.) ist das ESSzimmer; ↑BIron;
      THIS     is the dining room   ↑BIron;
02 B : schön; (0.2) □(0.1) das ist also
       nice                  this is then
   U:             |smile
       *img.1        *img.2
03 B : dein |esszimmer; |(1.0)
       your  dining room;
   U: ... |                  |turn |↓JA:,
                                    yes
                                   *img.3
04 U: (0.2) ↑STOP; (.) stop stop stop
                  *img.4
05 U: stop ↓stop
                  *img.5
```

Figure 11. Fragment 2 (Nov. VP03, 27:32 min)

So far, the system again performed as foreseen, and was able to engage in a typical three-part structure consisting of conversational offer (01), response (02–03) and ratification (03). To turn away bodily at this stage, as the user did in this fragment (img. 3), is a common procedure in natural interaction to signal the closure of the previous action and the preparation of a next activity (e.g., Schegloff, 1998). However, as the interaction continued, we can see that a misunderstanding occurred: when the user once again reoriented toward BIRON, she discovered that the robot was turning slowly away from her (img. 4). She treated this as an inappropriate action: She rushed to follow the robot (img. 5), yelling "stop" a couple of times (04–05) until she had managed to re-assume a face-to-face position with the robot and make it indeed come to a halt (img. 6).

This reveals that on the systems level, the robot's turning is part of the location-learning task, whereas this is not the case for the user. Having brought the previous sequence to closure officially and visibly ("yes" + turn away) and simultaneously started to project a next action, there is no relevant next for her to be expected from the robot at this stage. Therefore, she has to learn why the robot turns.

Observations like this point to the general problem of action structuring within interaction and the issue of how far we are able to equip robots with mechanisms that will allow users to interact more intuitively with them. For cases like this, we need to enable the robot to detect and interpret such user behavior as being meaningful on a structural level of interaction organization so that it can react accordingly by, for example, stopping its current activity or commenting on the pursuit of its action ("let me take a closer look at it"). Taking such aspects into account would, consequentially, lead to a more advanced monitoring of the user's actions and a breaking up of turn units in order to design robot systems that would be sensitive to the ongoing participation and engagement of the user (Kuzuoka et al., 2008).

13. Conclusion

In this paper, we have introduced our approach to evaluate the system and interaction levels of a human–robot interaction in order to achieve closed-loop design and implementation development. This approach tests components within a system and interaction context, and permits conclusions on how they work under realistic conditions and how they can be adapted to the challenges posed by the interaction and the complexity of the system. Our approach focuses on deviations from prototypical interaction scripts, or, in other words, on systematic problems that occur in the course of the interaction. Analyzing these allows us to systematically (a) assess the relevance of certain deviations, and (b) to treat them appropriately. SInA in particular facilitates an interdisciplinary analysis by integrating all levels of annotations—system internals and interaction course—into one annotated corpus that researchers from different disciplines can work on together.

Since SInA is based on TA, it enables researchers to analyze different tasks and systems with the same tool. Results can be compared easily with findings from analyses of other systems. Even though all annotations on the system level can be created automatically, a lot of work has to be put into the annotation of video recordings of user studies for any analysis on the interaction level. While this process is laborious, it is the basis for CA that allows very detailed descriptions of the interaction. However, we reduce this effort by first identifying the most relevant deviations using quantitative TA, following a hierarchical analysis scheme proceeding from task-oriented annotations to detailed CA.

A first case study is presented here. It shows that a high percentage of deviations can actually be allocated to concrete patterns. This can lead to specific improvements of the components themselves, of the robot feedback, and of the system architecture.

SInA is designed for use in long-term iterative development processes. Hence, its application is on-going work. Our future research will continue with iterations of user studies to incorporate the findings from such analyses and to gain a sufficiently large body of results on which to base long-term adaptations of the underlying interaction models. Moreover, the method will be used to compare our analysis to the analysis of other systems, tasks, and scenarios.

Acknowledgment

This work has been supported by CoR-Lab (Research Lab for Cognition and Robotics), the German Research Association (DFG) within the Collaborative Research Center 673, Alignment in Communication, and by the iTalk (Integration and Transfer of Action and Language Knowledge in Robots) project (EU ICT Cognitive Systems and Robotics Integrating Project FP7–214668).

Notes

[1] http://logging.apache.org/{log4cxx,log4j}

[2] www.lat-mpi.eu/tools/elan/

[3] Transcription Convention: The participants' talk is transcribed as it occurs, i.e., including pauses, hesitations, etc. (see Selting et al., 1998). Utterances are generally noted in lower case; upper case is used to mark stressed syllables (e.g., "BIron") and interpunctuation is used to signal specific prosodic features (e.g., in "↑Biron;" the prosodic contour starts with a high onset and falls at the end). The participants' visual conduct is annotated as short glosses, written in italics and generally accompanied by a frame grab of the video. The sequential relationship between events is rendered by their positioning on a virtual timeline.

References

Adams, J.A. (2005). Human–robot interaction design: Understanding user needs and requirements. *Proceedings of the 2005 Human Factors and Ergonomics Society 49th Annual Meeting,* Orlando, FL, USA, Cognitive Engineering and Decision Making, 447–451(5).

Burghart, C., Holzapfel, H., Haeussling R., & Breuer, S. (2007). Coding interaction patterns between human and receptionist robot. *Proceedings of Humanoids 2007*, Pittsburgh, PA, USA.

Christensen, H., & Topp, E. (2006). Topological modelling for human augmented mapping, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2257–2263.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics, 32*, 1489–1522.

Guivant, J., & Nebot, E. (2001). Optimization of the simultaneous localization and map-building algorithm for real-time implementation. *Transactions on Robotics and Automation, 17*, 242–257.

Hanheide, M., & Sagerer, G. (2008). Active memory-based interaction strategies for learning-enabling behaviors. *17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN),* Munich, Germany.

Hüttenrauch, H., Green, A., Oestreicher, L., Norman, M., & Severinson-Eklundh, K. (2004). Involving users in the design of a mobile office robot. *IEEE Transactions on Systems, Man and Cybernetics, Part C, 34(2).* May 2004, 113–124.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *The Journal of the Learning Sciences, 4*(1), 39–109.

Kim, H., & Kwon, D. (2004). Task modeling for intelligent service robot using hierarchical task analysis. *Proceedings of the 2004 FIRA Robot World Congress.*

Kuzuoka, H., Pitsch, K., Suzuki, Y., Kawagucchi, I., Yamazaki, K., Yamazaki, A., Kuno, Y., Luff, P., & Heath, C. (2008). Effect of pauses and restarts on achieving a state of mutual orientation between a human and a robot. *Proceedings CSCW 2008,* 201–204.

Li, S., Wrede, B., & Sagerer, G. (2006). A computational model of multi-modal grounding. *Proceedings ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006*, ACL Press, 153–160.

Lohse, M., Hanheide, M., Rohlfing, K.J., & Sagerer, G. (2009). Systemic Interaction Analysis (SInA) in HRI. *Proceedings of HRI conference 2009*, San Diego, CA, USA.

Muhl, C., Nagai, Y., & Sagerer, G. (2007). On constructing a communicative space in HRI. In *KI 2007: advances in artificial intelligence: 30th Annual German Conference on AI, Osnabrück, Germany: proceedings*. Joachim Hertzberg, Michael Beetz, Roman Englert (Eds.), Springer, 264–278.

Peltason, J., Siepmann, F.H., Spexard, T.P., Wrede, B., Hanheide, M., & Topp, E.A. (2009). Mixed-initiative in human augmented mapping. *Proceedings of ICRA 2009*, Kobe, Japan.

Sacks, H. (1992). *Lectures on conversation*. Malden, MA: Blackwell Publishers.

Schegloff, E.A. (1998). Body torque. *Social Research, 65*(3), 535–396.

Selting, M., Auer, P., Barden, B., Couper-Kuhlen, E., Günthner, S., Meier, C., Quasthoff, U.M., Schlobinski, P., & Uhmann, S. (1998). Gesprächsanalytisches Transkriptionssystem (GAT) [Conversation analytic transcription system]. *Linguistische Berichte, 173*, 91–122.

Stanton, N.A. (2006). Hierarchical task analysis: Developments, applications, and extensions. *Applied Ergonomics, 37*(1), 55–79.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Workshop on common metrics for human–robot interaction. *Human–Robot Interaction Conference*.

Yamazaki, A., Yamazaki, K., Kuno, Y., Burdelski, M., Kawashima, M., & Kuzuoka, H. (2008). Precision timing in human–robot interaction: Coordination of head movement and utterance. *Proceedings CHI 2008*, 131–140.

Yuan, F., Swadzba, A., Philippsen, R., Engin, O., Hanheide, M., & Wachsmuth, S. (2009). Laser-based navigation enhanced with 3D time of flight data. *Proceedings ICRA 2009*, Kobe, Japan.

Authors' Addresses

Manja Lohse, Karola Pitsch, Katharina Rohlfing, & Gerhard Sagerer
Applied Informatics, Bielefeld University
Universitätsstraße 25
33619 Bielefeld, Germany

{mlohse, kpitsch, rohlfing, sagerer}@techfak.uni-bielefeld.de

Marc Hanheide
Robotics and Cognitive Architectures Group School of Computer Science
University of Birmingham, B15 2TT, UK
m.hanheide@cs.bham.ac.uk

About the authors

Manja Lohse received her diploma in Applied Media Science from the Technical University Ilmenau, Germany in 2006. She is a Ph.D. student in the Applied Informatics Group at Bielefeld University, Germany. Since January 2009, she has been a member of the "Research Institute for Cognition and Robotics (CoR-Lab)" in the research group Hybrid Society. Her research focuses on evaluation and user expectations in human–robot interaction.

Marc Hanheide received a Diploma in Computer Science from University of Bielefeld, Germany, in 2001 and a Ph.D. degree (Dr.-Ing.) in Computer Science from the same university in 2006. He worked in the European Union projects VAMPIRE (2005) and COGNIRON (2008). From 2007 to 2009 he held the position of a senior researcher in the Applied Informatics Group responsible for several projects in cognitive interaction and robotics. Since 2009 he is a research fellow at University of Birmingham, UK, at the School of Computer Science affiliated with the European integrated project CogX. His fields of research include robotic architectures, human-robot interaction, multi-modal perception, and intelligent systems. He is affiliated with the Cluster of Excellence "Cognitive Interaction Technology (CITEC)". He is a member of IEEE.

Karola Pitsch received her Ph.D. in linguistics from Bielefeld University, Germany in 2006. After work as a postdoc researcher in the EU project PaperWorks (2005–2008) at King's College London, she joined the Applied Informatics Group at Bielefeld University for the EU-project iTalk in 2008. She is affiliated with the "Research Institute for Cognition and Robotics (CoR-Lab)." Her research focuses on multimodal aspects of social interaction and draws on conversation analysis. Research areas include workplace studies, CSCW, social learning, and (second) language acquisition. Recently, she has begun to apply insights from the study of human interaction to the design of human–robot interaction.

Katharina J. Rohlfing received her Master's degree in Linguistics, Philosophy, and Media Studies from the University of Paderborn in 1997. In 1999 she joined the Graduate Program Task-Oriented Communication at Bielefeld University and obtained her Ph.D. in Linguistics in 2002. She did postdoctoral work at San Diego State University, the University of Chicago, and Northwestern University. In 2006, she became a Dilthey Fellow (Volkswagen Foundation). Since 2008, she has been head of the Emergentist Semantics Group, Center of Excellence Cognitive Interaction Technology at the Bielefeld University. She is interested in the environmental cues supporting the development of meaning.

Gerhard Sagerer received his diploma, Ph.D. (Dr.-Ing.), and *venia legendi* (Habilitation) in computer science from the University of Erlangen-Nürnberg, Germany, in 1980, 1985, and 1990. He is now Professor of Informatics at Bielefeld University, Germany, and head of the research group for Applied Informatics. His fields of research include image and speech understanding, interaction, and the application of pattern understanding methods to natural science domains. He is on the management boards of the "Research Institute for Cognition and Robotics (CoR-Lab)" and the Cluster of Excellence "Cognitive Interaction Technology (CITEC)," and belongs to the German Computer Society (GI), EURASIP, and IEEE.