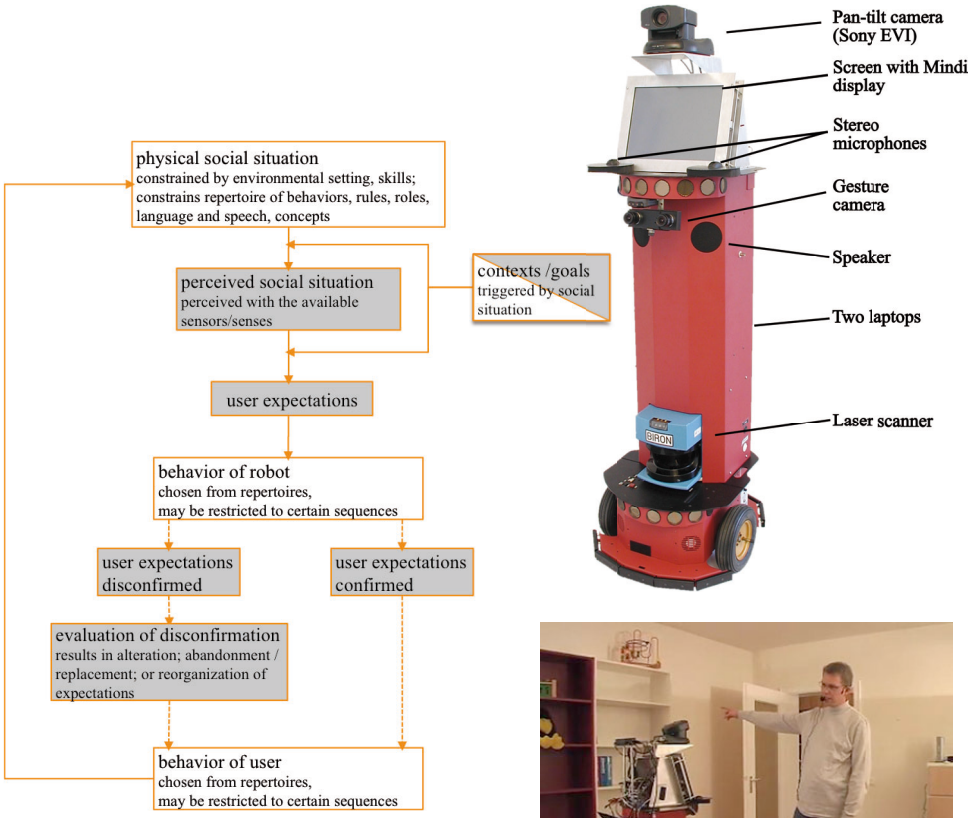
Investigating the influence of situations and expectations on user behavior - empirical analyses in human-robot interaction

Manja Lohse



Diplom Medienwissenschaftlerin Manja Lohse
CoR-Lab
Technische Fakultät
Universität Bielefeld
email: mlohse@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung des
akademischen Grades Doktorin der Naturwissenschaften (rer. nat.).
Der Technischen Fakultät der Universität Bielefeld
am 16.12.2009 vorgelegt von Manja Lohse,
am 19.04.2010 verteidigt und genehmigt.

Gutachter:
PD Dr. Katharina Rohlfing, Universität Bielefeld
Prof. Dr.-Ing. Martin Buss, Technische Universität München
PD Dr.-Ing. Sven Wachsmuth, Universität Bielefeld

Prüfungsausschuss:
Prof. Dr. Philipp Cimiano, Universität Bielefeld
PD Dr. Katharina Rohlfing, Universität Bielefeld
Prof. Dr.-Ing. Martin Buss, Technische Universität München
PD Dr.-Ing. Sven Wachsmuth, Universität Bielefeld
Dr.-Ing. Michael Pardowitz, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier nach ISO 9706

Investigating the influence of situations and expectations on user behavior - empirical analyses in human-robot interaction

Der Technischen Fakultät der Universität Bielefeld zur Erlangung des Grades

Doktor rerum naturalium

vorgelegt von

Manja Lohse

Bielefeld, Dezember 2009

Abstract

Social sciences are becoming increasingly important for robotics research as work goes on to enable service robots to interact with inexperienced users. This endeavor can only be successful if the robots learn to interpret the users' behavior reliably and, in turn, provide feedback for the users, which enables them to understand the robot.

In order to achieve this goal, the thesis introduces an approach to describe the interaction situation as a dynamic construct with different levels of specificity. The situation concept is the starting point for a model which aims to explain the users' behavior. The second important component of the model is the expectations of the users with respect to the robot. Both the situation and the expectations are shown to be the main determinants of the users' behaviors.

With this theoretical background in mind, the thesis examines interactions from a home tour scenario in which a human teaches a robot about rooms and objects within them. To analyze the human expectations and behaviors in this situation, two main novel methods have been developed. In particular, a quantitative method for the analysis of the users' behavior repertoires (speech, gesture, eye gaze, body orientation, etc.) is introduced. The approach focuses on the interaction level, which describes the interplay between the robot and the user. In the second novel method, also the system level is taken into account, which includes the robot components and their interplay. This method serves for a detailed task analysis and helps to identify problems that occur in the interaction.

By applying these methods, the thesis contributes to the identification of underlying expectations that allow future behavior of the users to be predicted in particular situations. Knowledge about the users' behavior repertoires serves as a cue for the robot about the state of the interaction and the task the users aim to accomplish. Therefore, it enables robot developers to adapt the interaction models of the components to the situation, actual user expectations, and behaviors. The work provides a deeper understanding of the role of expectations in human-robot interaction and contributes to the interaction and system design of interactive robots.

Acknowledgments

"Luck is when opportunity knocks and you answer" (Author unknown)

While being a PhD student I was lucky in many respects and that is why I want to thank everybody who contributed to that feeling. First of all, I was very lucky that Gerhard Sagerer gave me the chance to show that someone with a degree in Applied Media Science can contribute to robotics research.

Also, I want to very much thank my supervisor Katharina Rohlfing who was always open for questions. Katharina, thank you for your time, effort, and ideas that made writing this thesis a special learning experience for me. I'm also very much obliged to Martin Buss and Sven Wachsmuth who agreed to review the thesis even though it does not have an engineering background.

This openness towards other disciplines is also something I feel really lucky to have encountered in the Applied Informatics. Thanks to all my colleagues for making the last years of working in Bielefeld a really good experience. I especially want to thank Angelika Dierker who I liked to share office with and Marc Hanheide. Marc, it's invaluable to find people whose thoughts combine with one's own to so many good ideas, a lot of which being present in this thesis.

I also want to thank all the people that were involved in the studies that are the basis for this thesis, especially Christian Lang who conducted one of the object-teaching studies, all my colleagues who tried to make BIRON smart enough to interact with the users, and Lisa Bendig and Sascha Hinte who accompanied my research as student assistants.

I thank my parents for supporting me in everything I do and helping me to find my way by questioning the important decisions in my life. Finally, there is one thing that I want to say in German. René, ich danke dir für all die Geduld und Unterstützung, die du mir gegeben hast, obwohl du so manche meiner Launen sicher nicht verstehen konntest. Ich bin sehr glücklich, dass ich auch die kommenden Herausforderungen mit dir zusammen erleben darf!

Table of content I

Table of content II

List of figures

List of tables

1 Introduction

Imagine that you have ordered a robot to assist you in the household. Today it has been delivered and is now waiting in your living room ready for operation. For you, this situation is completely new because you have never interacted with such a robot before. Even though you have read the specifications of the robot on the company website, you are not quite sure how to operate it. All you know is that you can talk to the robot like you would talk to a human assistant and that it has to learn about your home before it can solve tasks for you. But how would you try to teach these things to the robot? Would you talk to the robot like you would with an adult? Would you treat it like a child or a pet? And what behaviors would you expect from the robot?

Situations in which novice users come into contact with service robots that operate in close proximity to them and share the same spaces are becoming more and more common, because robots are being developed to enter these spaces in order to enrich and ease people's lives (IFR, 2007). Most of the 3.5 million service robots sold to private users up to 2006 were vacuum cleaners, lawn-mowing robots, and a wide variety of entertainment and leisure robots. The IFR (2007) has estimated that another 3.5 million personal service robots would be sold between 2007 and 2010. Thus, millions of people are encountering first-contact situations with robots and need to find out how to operate them. Since the systems that are sold to private users to date have been rather simple they do not require special training. For the future, much more complex scenarios such as the assistant for the home are envisioned. However, these will depend on easy operation and a high level of system-human integration (Engelhardt & Edwards, 1992). This can only be achieved if the developers know *how* the users interact with the robots, what expectations they have, and how these change during the interaction. These issues are addressed here in order to provide thorough descriptions of users' behaviors and their expectations to the designers of human-robot interactions.

1.1 Human-robot interaction (HRI) – definition and usage

In the following, three definitions of human-robot interaction (HRI) are introduced that serve as a starting point for the discussion of existing perspectives on the field. The first definition by Fong, Thorpe, and Baur (2001) is the most general one:

> *"Human-robot interaction (HRI) can be defined as the study of humans, robots, and the ways they influence each other." (Fong, Thorpe, & Baur, 2001, p.11)*

The authors focus on the subjects of interest, humans and robots, and stress that HRI is interested in how they interact, or "how they influence each other". This definition is very broad and does not specify research disciplines that are involved in HRI as does the next definition by Wagner and Arkin (2006):

> *"Human-robot interaction (HRI) is a subfield of AI that combines aspects of robotics, human factors, human computer interaction, and cognitive science […]. The details of how and why humans and robots interact are focal research areas within HRI […]. Typically, HRI research explores mechanisms for interaction, such as gaze following, smooth pursuit, face detection, and affect characterization." (Wagner & Arkin, 2006, p.291)*

Apart from naming the disciplines included in HRI, this definition also specifies the robot and the human as subjects of interest that are analyzed in how they interact. Moreover, the definition raises the question of "why" they interact and introduces some examples for interaction mechanisms that researchers are interested in. Finally, a third definition by Goodrich and Schultz (2007) shall be mentioned:

> *"Human-Robot Interaction (HRI) is a field of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans. Interaction, by definition, requires communication between robots and humans." (Goodrich & Schultz, 2007, p.204)*

Goodrich and Schultz (2007) stress the processes of understanding, designing, and evaluating robots. Their perspective is inspired by usability research. The interaction does not only have to be analyzed as implied by the first definitions, but the aim of HRI is to actively improve the robot for it to be useful to humans. Similar to the other definitions, the authors also point out that humans and robots take part in the interaction process.
All definitions agree on the most general assumption that HRI includes humans as well as robots. Apart from this, the questions raised in the definitions differ and, thus, imply various aims of HRI. The aims are influenced by the perspective that researchers have on a field. That is why in the following some perspectives on HRI are introduced.
Kiesler[1] stated at the panel discussion on the question "What is HRI?" at the 2007 HRI conference in Amsterdam that "HRI is not a discipline but an area of research". This statement evokes one question: What constitutes a discipline? Siegrist sums up four characteristics of scientific disciplines (Siegrist, 2005, p.7). Every discipline has:

1. a certain terminology (Begriffssprache)
2. specific methods or research techniques (Forschungstechniken)
3. an increasing amount of descriptive information regarding relevant findings within the discipline (Beobachtungswissen)
4. to a certain extent established theories that help to explain, predict, and alter interrelations between phenomena (Theorien)

[1]Hillman Professor of Computer Science and Human-Computer Interaction at CMU (http://www.cs.cmu.edu/~kiesler/index.html, 12.10.2009)

In contrast to these characteristics, so far, there is not yet a general accepted corpus of knowledge, set of scientific questions, methods, or strategies in HRI research. Goodrich[2] described this situation similarly at the same panel. He stated that HRI is just now emerging as a field and that there are no strictly defined standards regarding how to conduct research and how to present results in papers. According to Scasselatti[3] (at the "Social Interaction with Intelligent Indoor Robots [SI3R]" Workshop, ICRA'08, Pasadena, Ca, USA), when designing a new computer display, textbooks about ergonomics and other resources can be consulted, but nothing comparable exists for social robotics. In HRI, researchers draw from case studies, specifics about certain areas, and models from other fields. In accordance with these statements, it is noticeable that the first journal in the field (International Journal of Social Robotics[4]) has been launched only recently. The first edition appeared in January 2009.
These statements and the definitions introduced above lead to the conclusion that not one unified view, but different perspectives on HRI exist. In the following, some of these will be introduced based on different interaction situations and approaches to the field by researchers with different scientific backgrounds.

Takeda, Kobayashi, Matsubara, and Nishida (1997) distinguish three forms of HRI:

- intimate (multimodal direct interaction, robots can use their bodies for communication, people and robots are spatially close to each other)
- loose (people and robots are in different locations)
- cooperative (robots cooperate with each other to exploit all functions they have together)

The differentiation is based on the flow of information and control. On this basis, Thrun (2004) distinguishes indirect and direct interaction:

> *"In indirect interaction, the operator controls the robot, which communicates back to the operator information about its environment and its task. In direct interaction, the information flow is bi-directional: information is communicated between the robot and people in both directions, and the robot and the person are interacting on equal footing." (Thrun, 2004, p.15)*

Accordingly, in indirect interaction, the human controls the robot and direct interaction is bi-directional at all times. Next to this differentiation also the way one looks at HRI brings about different approaches. Dautenhahn (2007) distinguishes:

[2]Associate Professor at the Computer Science Department of the Brigham Young University (http://faculty.cs.byu.edu/~mike/, 13.10.2009)
[3]Associate Professor, Department of Computer Science, Yale University (http://cs-www.cs.yale.edu/homes/scaz/, 13.10.2009)
[4]International Journal of Social Robotics (http://www.springer.com/engineering/journal/12369, 10.11.2009)

- robot-centered HRI: the robot is seen as a creature with its own goals based on its motivations, drives and emotions; the interaction with people serves to fulfill some of its needs
- human-centered HRI: the robot as a system fulfills its task in an acceptable and comfortable manner from the point of view of the human
- robot-cognition-centered HRI: this approach emphasizes the robot as an intelligent system in a traditional AI sense, the robot has to have cognitive abilities (for example, for learning and problem solving)

The four differentiations by Takeda et al. (1997), Thrun (2004), and Dautenhahn (2007) serve as an introduction to different forms that HRI can take and different viewpoints one can choose in order to look at the field. To add to this, now some views on HRI will be outlined which are based on expert opinions discussed at "NEW HRI, an ICRA 2008 Workshop on Unifying Characteristics of Research in Human-Robot Interaction". At this workshop, Forlizzi[5] postulated to look at HRI problems holistically, meaning that attention has to be paid to the context and the users. Developments should be based on observations of human needs. Thus, iterative studies conducted in the field and not in the laboratory are a key part of the work, in order not to develop technology just for its own sake but to help people. Matarić[6] supported the idea that the developers must think about the user. They must work with the users from the beginning. Again she emphasized that user studies should not be conducted in the laboratory. In her opinion, researchers should build human-centered technology to assist people. Also Christensen[7] fostered the idea to design robots for interaction with humans and to evaluate usability with users in short-term as well as long-term studies. All three researchers stress the importance of human-centered HRI which also guides the following analyses.

Moreover, to succeed in the field a large amount of *interdisciplinarity* is indispensable. This is underlined by the following sentence from the "Welcome message" in the 2008 HRI conference proceedings (1): "Human-robot interaction is inherently inter-disciplinary and multi-disciplinary". The common goal of "systematizing the fundamental principles of social interaction and evaluating the emerging relationships between humans and robots" (Sabanovic, Michalowski, & Caporael, 2007) is what brings the researchers together. Of course this objective is rather abstract and usually researchers work on much more concrete goals. From a technical perspective, the goal of HRI is to build robots that get along in the context and the environment they are used in. Single functions, which support this goal, are evaluated. From a usability point of view, the components not only have to operate as the developer conceptualizes them, meaning that they fulfill their functions and are technically stable, the system also has to be easy and safe to use, and socially acceptable (for example, Dix, Finlay, Abowd, & Beale,

[5]Associate Professor at the School of Design and Associate Professor at the Human-Computer Interaction Institute, and the A. Nico Habermannn Chair in the School of Computer Science at Carnegie Mellon University (CMU) (http://www.design.cmu.edu/show_person.php?t=f&id=JodiForlizzi, 11.10.2009)

[6]Professor in the Computer Science Department and Neuroscience Program at the University of Southern California (USC), founding director of USC's interdisciplinary Center for Robotics and Embedded Systems (CRES) and co-director of the USC Robotics Research Lab (http://www-robotics.usc.edu/~maja/, 11.10.2009)

[7]KUKA Chair of Robotics at the College of Computing Georgia Institute of Technology. He is also the director of the Center for Robotics and Intelligent Machines at Georgia Tech (http://www.cc.gatech.edu/~hic/Georgia-HomePage/Home.html, 11.10.2009)

2004; Nielsen, 1993; Shneiderman, 2002). Since humans are social beings we can assume that the usability of robots improves if they also have social capabilities. Social-emotional intelligence helps humans to understand others and to interact with them (Breazeal, 2003). Therefore, one more goal researchers are focusing on is building *social robots* because "developing an intelligent robot means developing first a socially intelligent robot" (Dautenhahn, 2007). Social robots are "those that people apply a social model to in order to interact with and to understand" (Breazeal, 2003, p.178).

All the aspects discussed in this section lead to the usage of the term in the following. Here HRI describes the interaction of one user with one robot in an intimate and direct form. Accordingly, both interactors share the same physical space. The robot is social because it interacts with means that the users know from human-human interaction (HHI) and a social model is applied to it. The aim of the HRI research is the human-centered adaptation of the robot system. This goal is inherently interdisciplinary. To achieve it, the systems need to be tested in user studies that take place in realistic environments and situations.

1.2 Scenario and robot system

The social robot BIRON (BIelefeld RObot CompanioN) is in the center of the user studies presented in the following. BIRON has been developed for the home tour scenario which focuses on multi-modal HRI to enable the robot to learn about a domestic environment and its artifacts, the appearance and location of objects, and their spatial temporal relations. In other words, the robot is guided through an apartment by the user and learns about rooms and objects. This is necessary because it is not possible to pre-program the robot for every possible environment. Hence, it has to learn about its surroundings with the help of the user. Once the robot has acquired the knowledge about its surroundings, it can serve as a kind of "butler" providing personal services (for example, laying the table, cleaning rooms). In the current implementation, the learning is the most important part of the scenario. Because of motor limitations the robot is not yet able to conduct any housework tasks. However, the scenario serves to explain to the users why the robot has to acquire knowledge about its surroundings.

In the home tour scenario, users are likely to be novices or advanced beginners, meaning that they have little or no experience interacting with a personal service robot for the home.[8] In fact, in the user studies presented in the following, most of the trials were first-contact situations,

[8]Hackos and Redish (1998) distinguish between four groups of users: novice, advanced beginner, competent performer, and expert. *Novices* are very goal and task oriented. They usually want to start the interaction right away without having to learn a lot about the system. According to Hatano and Inagaki (1992), *novices* depend on devices and materials, other people's knowledge, and knowledge provided in an externalized symbolic form. Their knowledge is highly situated, i.e., as soon as some situational factor changes, novices might not be able to solve the task anymore. They might also only be able to solve it with the provided aids. *Advanced beginners* focus on getting a job done quickly with the lowest effort possible. They are content using a few tasks out of the ones available. Advanced beginners begin to develop an empirically based mental model of the system. *Competent performers* use more tasks than advanced beginners. They have learned a sufficient number of tasks and formed a sound mental model of the system, which enables them to create a plan of action to perform the task, to recognize, diagnose, and solve problems. Finally, *expert performers* have a comprehensive and consistent model of the product functionality and the interface. They understand complex problems and solve them. Moreover, they are interested in learning about concepts and theories behind the product design (Hackos and Redish, 1998).

such that the users had never interacted with a social robot before. This is because this type of robot is not commercially available yet. However, the natural interaction with the robot is a means to quickly learn how to complete tasks efficiently. According to Maas (2007, p.6), "ideal" robots in the home should be able to communicate naturally and to a certain extent solve problems autonomously. Since this is one inherent goal in the home tour scenario, it is neither necessary, nor desirable that users of a home tour robot become experts. However, interaction with a robot at home is usually a long-term interaction. The scenario envisions that a user buys a robot and uses it as a companion for a long time. Therefore, over time the users develop mental models of the system when interacting with it and, automatically, learn about it. However, in the envisioned case they learn more about the interaction than about technical details.

In this context, only potential users provide us with enough data to develop robots that they enjoy interacting with. Therefore, this thesis is strongly anchored in the tradition of usability research and focuses on human-centered HRI. User expectations are in the center of it. Hence, a vital part is to learn about the human interaction partner. However, the following analysis does not aim to provide an in-depth description of human psychology in HRI but rather of the situations in the scenario. Some of the pre-defined situational factors that characterize the home tour are that the interaction between the user and the robot is intimate and direct. Both interaction partners are present in the same room and interact in real-time. The robot uses natural interaction modalities such as speech.

A basic ability the robot needs in the interaction with a user is person perception. Thus, it has to know whether a user is present and is willing to interact with it. Since one major task is to show objects to the robot, it has to interpret gestures, for example, pointing gestures to locate the position of the objects. Moreover, it must recognize the objects and the rooms that they are located in. Finally, it has to move within these rooms. The abilities of our robot with this respect will be described shortly.

Next to the scenario, also the appearance and functionality of the robot platform have a major influence on HRI. Due to its embodiment, the robot shares physical space with its user. The mobility allows it to approach the human instead of waiting to be addressed. This ability distinguishes the robot from virtual agents and computers. For the home tour scenario a domestic robot has to be employed that Young, Hawkins, Sharlin, and Igarashi (2008) define "to be a machine that (a) is designed to work with individuals and groups in their personal and public spaces, (b) has a dynamic spatial presence in those spaces, and (c) can "intelligently" interpret its environment and interact physically with it" (p.99).

BIRON was developed to have these abilities. It is based on an ActiveMedia™ Pioneer PeopleBot platform. The robot is equipped with a pan-tilt color camera at a height of 141 cm to acquire images of human faces. The camera is also used to show attention by looking in a certain direction and to actively explore the environment visually.

Below the camera, there is a screen which can be used to display the behavioral states of the robot with an animated character called Mindi (see Section 5.2.2). The character resembles BIRON. The Mindi pictures are directly related to the robot state, for example, if the robot is in the follow state, an animation is shown in which the Mindi has a happy face and is walking; if the user is not perceived well, the face of the Mindi character shows an apologetic expression

and a speech bubble appears next to the Mindi in which a picture of a fuzzy stick figure is displayed.
A pair of AKG far-field microphones is located right below the display at a height of approximately 106 cm. With their help, BIRON can localize speakers. For speech recognition itself, the user wears a headset. Moreover, below the microphones there is an extra camera which is used to recognize gestures, and two speakers for speech output. A SICK laser range finder, mounted at a height of 30 cm, measures distances within a scene.
BIRON is able to follow a person and to autonomously move around. Additionally, it can track people and selectively pay attention to humans looking at it. Further technical information about BIRON is provided in Hanheide and Sagerer (2008).

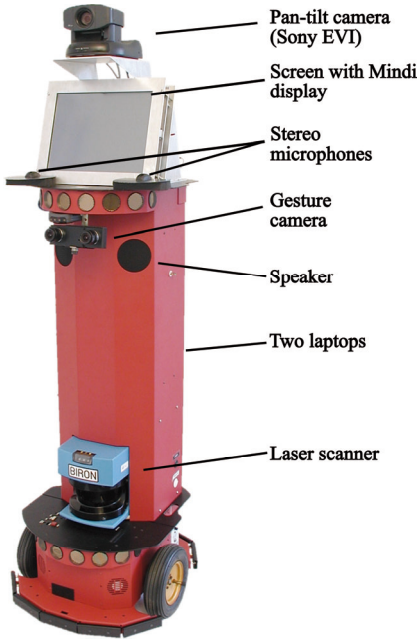


Figure 1-1. BIRON (BIelefeld RObot companioN)

1.3 Contribution

This scenario and the robot raise many research questions. Recalling the situation that was introduced at the beginning of this introduction, the questions that it implies focus on the expectations of the users and the change of these expectations in different situations. So far, little research exists regarding the influence of users' expectations and the situations on HRI and little is known about the impact of both concepts on the interaction, even though in HHI they have been found to crucially influence the behavior of human participants in various experiments (see Chapter 2).
Some researchers in HRI have looked at expectations that the users have before the interaction (see Section 2.2.6). However, their approaches were often insufficient to predict the user behavior during the interaction and did not tell how the robot could be designed to improve the interaction. That is why a model is needed that describes the dynamic changes of the expectations in the interaction situation and helps to better understand and predict the behavior of the users. Therefore, the first aim of this thesis is to develop such a model to show the dynamic relationships between expectations, situations, and HRI (see Section 2.3).

With respect to the model, it is necessary to introduce an approach to systematically describe the situation as such. The description has to include a thorough analysis of the behavior of the users. This is what mainly influences the interaction because the behavior of the robot is deterministic. To develop an approach for such a description is the second aim of this thesis (see Section 2.1.1). In order to fully understand the user behavior, it needs to be analyzed qualitatively and quantitatively strongly considering its multimodality. As was pointed out above, the field also is in need of more methods to achieve this goal and to systematically research questions that come up in the context of HRI in general. Therefore, the third aim is to introduce qualitative and quantitative methods that combine technical and social insights to research specific questions in the context of HRI. In order to be valuable contributions to the field, the methods aim to be generalizable, applicable to many scenarios and robots, efficient, and easy to use. Moreover, they need to account for the multimodality of the interaction situation. Chapter 3 presents the methods that were developed with these aspects in mind.
Such a model and methodology can also advance the literature about expectations by insights that are not based on HHI but on HRI which is the fourth aim of this thesis. It intends to show that HRI is valuable to support research on expectations because it profits from the fact that human behavior can be analyzed in relation to robot behavior which can be defined exactly. In other words, the advantage of working with robots is that they can display the same behavior over and over again and, thus, always influence the situation in a similar way. Therefore, changes in user behavior can be attributed to the users' perception of the situation alone. Thus, new insights about the expectations of the users can be gained which the research on expectations in general can benefit from.

1.4 HRI user studies

Three user studies were conducted with the robot BIRON to address the aims introduced above. In each study, the robot displayed different skills and behaviors and the users had to complete different tasks. In the following, the studies are described.

1.4.1 Object-teaching studies

The object-teaching studies focused on the task of teaching objects to the robot. The first study was designed to evaluate how and when users change their discursive behavior when interacting with a robot in an object-teaching task (Lohse, Rohlfing, Wrede, & Sagerer, 2008). It investigated which verbal and nonverbal behaviors the users applied. Verbal behaviors included utterances of the users and nonverbal behaviors focused on gestures of the users.
In the study, BIRON was limited to a few communicative capabilities. Next to general expressions like "Hello", the robot used a small set of feedback utterances regarding the task (the utterances have been translated by the author because the study was run in German):

- "That's interesting. I really like it."
- "Yes please?"
- "I didn't completely understand you. Could you please repeat that?"
- "Pardon."

- “Sorry, I’m still young and can’t do this.”
- “Sorry, I can’t search for objects right now.”
- “Sorry, I don’t know.”

This restricted set of answers, along with the concrete object-teaching task, allowed the behaviors between users to be compared. It also led to numerous miscommunication situations that triggered behavioral changes. The hypothesis now was that the participants change their discursive behavior (speech and gesture) to solve the miscommunication situations, or, more generally, the participants would adapt to the feedback of the robot. Figure 1-2 depicts the setup of the study.

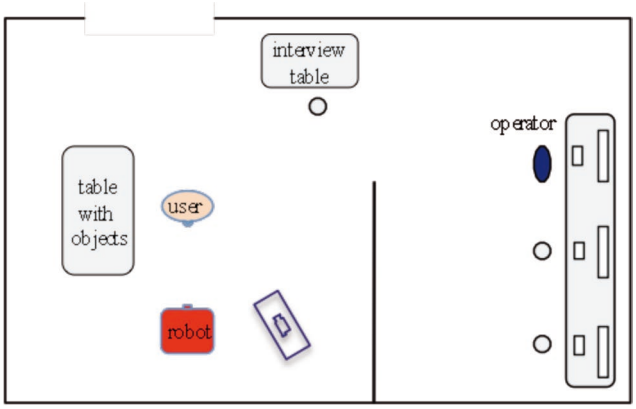


Figure 1-2. Setup of the object-teaching study

For the analysis of this first study, coding schemes for object-teaching tasks were developed which contained all behaviors that the users showed in their verbal behavior and their gestures (see Section 4.1). The coding schemes are here applied and adapted in a second study.

This second study was conducted as a follow-up to the first object-teaching study. The participants again should teach the names of several manipulable everyday objects to the robot and validate that it had learned the objects. Once more, it was not specified how they should term the objects and how they should present them (for example, pointing to them or lifting them up). Again, the robot did not drive around but reacted to the subjects by speech production and movements of its pan-tilt camera.

While in the first study the robot acted autonomously and there was no control over the interaction and the reactions that would be evoked by it, the second study was a Wizard of Oz study, meaning that the robot was teleoperated which the participants did not know. A script predefined how the robot reacted to a certain object which was shown at a certain point of time during the interaction. Therefore, one could be sure whether or not the answer was correct and the users’ behaviors could be compared despite the greater amount of robot utterances. Every user completed two trials. In one of them the robot recognized a lot of objects correctly and in the other one it failed repeatedly.

Compared to the first study, the data of the second study bear some considerable advantages for the analyses presented in the following. They can be readily analyzed comparing the positive and the negative interaction situations to determine whether and how the course of the inter-

action influences the users' behavior. The SALEM (Statistical Analysis of Elan files with Matlab) method has been developed to conduct this kind of comparative analysis (see Section 3.2).

Moreover, the Wizard of Oz setting increased the number of robot behaviors by a considerable degree. Altogether, more reactions of the users to more situations could be analyzed with the question in mind of whether the behaviors that were identified in the first study would still be adequate to describe the interaction. The analysis also included human gaze as another modality that provides insights into the course of the interaction. Finally, the interplay between the modalities could be analyzed to receive a more comprehensive picture of HRI in an object-teaching task. Even though the first study was not analyzed in such depth, it played an important role in the study design and the development of the coding schemes.

1.4.2 Home tour studies

While the object-teaching studies are valuable to research the object-teaching situation, the home tour can only be studied in a realistic environment because it contains more tasks such as teaching rooms and guiding the robot. Also Green (2009) found it highly important in the context of the home tour scenario to conduct studies in real apartments. Young et al. (2009) support this claim by stating that the subjective consumer perceptions of what robots are, how they work, and what they are capable of doing or not in a domestic environment can only be understood in the context of social interactions. These take place within the scenarios. Another argument in favor of scenario-guided testing is that design decisions have to be based on data and should not rely on assumptions (Hackos & Redish, 1998, p.50). Hence, conducting the studies in realistic environments is crucial because situations cannot be simulated in the laboratory. That is why BIRON was evaluated in a real apartment (see Figure 1-3).

This has turned out to be positive in the sense that the environment makes it easier for the subjects to understand the scenario. Moreover, the environment is valuable in that the users develop realistic expectations towards the robot because the surrounding causes restrictions to the system that would also occur in real user homes; for example, changing lightening conditions, narrow doors, and carpets impairing the robot's movements. The realistic test environment enables the researchers to adapt the robot to these problems.

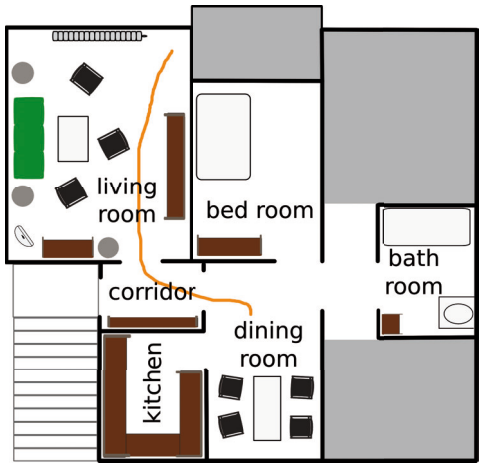


Figure 1-3. Layout of the apartment and the path the robot had to be guided

1.5 Outline of the thesis

The data acquired in the user studies are analyzed with respect to the impact of situation and expectations. These two basic theoretical concepts that guide the following analysis are introduced in Chapter 2. It shows that the situation can be analyzed on different levels of specificity and identifies HRI as a social situation. Moreover, it describes the influence of expectations on the situation. The theoretical overview results in assumptions concerning expectations of the users that can be drawn for HRI and in a model that describes the relation between situation, expectation, and user behavior in HRI.

Chapter 3 presents the methods that were developed for the purpose of researching the HRI situation and the assumptions connected to the expectations of the users. The methods include both qualitative and quantitative approaches for the analysis of single behaviors of the users, the structure of the interaction, and the users' reflections about the interaction.

In Chapter 4, some of the methods are applied to the object-teaching studies that were conducted in the laboratory. The chapter presents coding schemes for user behavior in object-teaching situations and the quantitative SALEM (Statistical AnaLysis of Elan files in Matlab) of speech, gesture, and gaze in two situations that differ with respect to the success of the task completion. Moreover, it discusses the effects of different phases of the interaction on the behavior of the user.

Chapter 5 focuses on the SALEM of the modalities gesture, gaze, and body orientation in the home tour studies. All the modalities are analyzed because they can help the robot to identify what the user is trying to do. The SInA (Systemic Interaction Analysis) method, which is applied to analyze the home tour tasks, will show how the behaviors differ between the tasks.

The thesis closes in Chapter 6 with a summary of the theoretical and methodological achievements, a discussion of the shortcomings, and future perspectives of the work.

2 Theoretical background for the introduction of the notions situation and expectation to HRI

In this chapter, the notions situation and expectation will be put on theoretical grounds in order to develop a model of situation and expectation for HRI. In the first part of the chapter, the notions of situation and context are contrasted in order to clarify how they are understood here and why a distinction is necessary (Section 2.1). Thereafter, expectation theory is introduced to illustrate what effects expectations have on interaction in general and on HRI in particular (Section 2.2). Finally, based on this theoretic evaluation, a situation- and expectation-based model of HRI is developed (Section 2.3).

2.1 Situation and Context

The terms *situation* and *context* are often used synonymously (Rohlfing, Rehm, & Goecke, 2003; Bierbrauer, 2005). However, they need to be separated because both concepts have their own relevance: while the situation is represented physically and can be perceived by actors, the context represents internal knowledge of the actors that is needed to handle the situations (Rohlfing, Rehm, & Goecke, 2003). In the following, both concepts are applied to HRI.

2.1.1 The concept of situation in HRI

We experience the world around us through a great many situations that each of us encounters everyday. We meet a friend on the bus, attend a speech at a conference, or interact with a robot. All these experiences, bound to certain situations, shape our picture of the world. But what is the physical situation that exists around us and what is a situation once we perceive it? These questions are addressed in the following.

2.1.1.1 The physical situation

According to Rohlfing, Rehm, and Goecke (2003),

> *"A situation consists of the spatiotemporal ordering of objects and agents alongside physically given constraints or characteristics like gravitational forces or light intensity." (Rohlfing, Rehm, & Goecke, 2003, p.133)*

The definition highlights the importance of space and time for the characterization of a situation. Moreover, situations include objects, agents and their actions. These actions are constrained physically. Magnusson's (1981a) definition is along the same lines. He distinguishes three kinds of environment: physical geographical (room with objects, etc.), biological (people in the room with their age, sex, etc.), and socio-cultural (rules, norms, and behavior based on the other two kinds of environments). Physical situations cannot only be analyzed by their content but also by their structure. According to Magnusson (1981a), sub-units of situations can be defined, for example, *stimuli* that function as signals in themselves (sounds, etc.) and *events* that can be described in terms of cause and effect.

In Craik's (1981) view, next to the components mentioned above, situations also include actions and cognitive aspects. That is where the actors come into play that perceive the situations and act within them. Even though the physical situations exist without anyone being present, they are of interest here when *perceived* by actors and framing their actions.

2.1.1.2 The perceived situation

The perceived situation has been defined by Magnusson (1981b) as follows:

> *"A perceived situation is defined here as an actual situation as it is perceived, interpreted, and assigned meaning or, in other words, as it is construed by and represented in the mind of a participant." (Magnusson, 1981b, p.15)*

Accordingly, the perceived situation depends on the cognitive processes that the person applies to interpret the physical situation. These processes differ between people. To refer to the example of attending a speech at a conference, certainly the listener interprets the situation differently than the speaker. But also another listener who might know the speaker or have more or less knowledge about the topic will perceive the physical situation differently. Throughout the situation, the physical information that is available to interpret it can be separated into three categories: a) information in the focus of attention, b) peripheral information which is considered but not attended to, and c) information that is not considered at all (Magnusson, 1981b). What is in the focus of attention changes continually. It depends on the situation and its conceptualization and also on adjacent situations that influence each other. Because of these interrelations no two situations are exactly the same (Rotter, 1981). Perceived situations change because time goes by and people make experiences. What has happened in a past situation changes a new situation. Because time has such a big influence on situations, researchers speak about a *dynamic* concept in contrast to personality traits which are a *stable* concept (for example, Rohlfing, Rehm, & Goecke, 2003, Bierbrauer, 2005). According to Smith (2005), dynamic systems are coupled to a physical world that is also dynamic. Therefore, the systems themselves change based on their history, very much driven by the multimodal nature of experience, i.e., the different senses offering different takes of the world, which are time-locked with each other. For this reason, modalities alter each other and dynamic concepts remain changeable. Based on this modular view, one could wonder why complex systems act coherently as Smith (2005) has pointed out. Coherence "is generated solely in the relationships between the components and the constraints and opportunities offered by the environment" (Smith, 2005, p.278). In other words, the environment (i.e., the situation) restricts the complex systems in such a way that they act coherently because it only offers a limited set of options. Since in the situation complex systems act coherently, situations have a certain degree of similarity and we are able to identify similar situations. This is the prerequisite for learning from one situation for another. Based on the similarity between situations we can make better-than-chance predictions of what is going to happen.

In 1973, Darley and Batson conducted an influential experiment which has shown the impact of small changes in situations on participants' behavior. The researchers told seminarians to give a

talk at another building on campus. One group was to hurry up on the way to the building where the talk had to be given; the other group was told that they had plenty time. On the way, the subjects encountered a person who obviously needed help. The results showed that people who had plenty time helped much more often than people who had been asked to hurry up (63% vs. 10 %). In the experiment, the change in situation affected the behavior of the subjects much more than their personality traits. They *construed* the situation differently.

Libermann, Samuels and Ross (2004) conducted another study showing that situations can be perceived quite differently. In the study, students played the same game under a different name (Community Game, Wall Street Game). The researchers found that students in the second condition (Wall Street Game) displayed significantly more competitiveness independent of how their competitiveness was rated beforehand. The students obviously construed the situation differently. Their awareness of the situation was affected by the name of the game and different expectations were raised.

Many more experiments have shown how situational factors can superpose personality in HHI. Other influential examples that support the hypothesis are the Stanford Prison experiment (Zimbardo, 2007) and the Milgram experiment (Milgram, 1974). Therefore, it is expected here that the situation as such is an important factor in interaction that also influences HRI.

It can be assumed that in all these experiments the situation factors outweighed the personality factors because many subjects construed the situation in the same way. Nevertheless, the construction process is an individual process. This has first been postulated by the *Gestalt psychologists.* They proposed that each human perceives the environment individually (see Aronson, Wilson, & Akert, 2003). Lewin (1935) applied this assumption to social perception, i.e., on the question of how people perceive other people, their motives, intentions, and behaviors. He was the first scientist who has recognized how important it is to take the perspective of the individuals to find out how they construe the environment (Aronson, Wilson, & Akert, 2003).

According to Lewin (1935), behavior is a function of the person and the environment. This means that, in general, actor personality and situations that the actor has encountered throughout life influence the actor's behavior; or in other words, nature as well as nurture interact to shape each person. However, often the question of whether person or situation dominates cannot be answered because influences of person and situation cannot easily be separated from each other (Heckhausen & Heckhausen, 2006). The influence on behavior depends on the sample of situations and persons. However, if many people in the same situation do the same thing, it can be concluded that the situation – and not the personality – is the main determinant, as was the case in the experiments introduced above.

However, the perception of the situation is also guided to a certain extent by the actors' personalities and interaction goals. In fact, lay people are intuitively prone to attributing behavior to personality and not to situations. While this might sometimes be reasonable, Ross and Nisbett (1993) report that people tend to overestimate the influence individual differences have on behavior. At the same time they underestimate that each person construes situations individually. This phenomenon is called *fundamental attribution error*. It is characterized by "the tendency to overestimate the extent to which people's behavior is due to internal,

dispositional factors and to underestimate the role of situational factors" (Aronson, Wilson, & Akert, 2003, p.13). Jones (1990) names determinants for why the fundamental attribution error occurs:

a. action is more dynamic and salient than context,
b. people have a richer vocabulary for personality since situational factors are taken for granted,
c. people are socialized to accept the act of others at face value,
d. inaccurate dispositional attributions are seldom questioned because they hardly ever disrupt the interaction process and they do not have negative consequences.

Ross and Nisbett (1991) contribute to these determinants by proposing that what people attend to, is what they attribute to. The authors refer to a study in which Arkin and Duvall (1975) found that behavior was attributed less to the environment when the environment was stable than when it was in motion. Moreover, Taylor and Fiske (1975) showed that if two people are in a situation and one can be seen better than the other one by an observer, than more causal attributions are made to the person who can be seen better.

Another important differentiation concerns attributions of an observer and of actors in the situation. Ross and Nisbett (1991) report that observers attribute behavior more strongly to the actors while the actors themselves tend to attribute their behavior to the situation. However, there is some evidence that when actors see themselves on videotape they make attributions similar to those of the observers. This leads to the assumption that attribution is guided by the focus of attention and actors and observers typically attend to different things.

Even though attributions are often made to the traits of actors, according to Ross and Nisbett, (1991), Asch, based on a series of experiments, claims that "people's responses to an object are often less reflective of their long-held attitudes and values than of the way they happen to construe the 'object of judgment' on a given occasion" (Ross & Nisbett, 1991, p.69). Thus, the responses do not reflect attitudes but changes in the situation. In experiments often the 'object of judgment' changes with the situation and not the 'judgment of objects'. In other words, attributions change because the situation and how the person construes it has changed; however, the personality stays rather stable. Thus, we can often learn much more about construction processes than about personality. The personality, however, influences choices that people make regarding situations they encounter. Which situations people enter, reveals what kind of person they are, or in other words, the fact that people spend their time in situations is a function of their personalities (Argyle, Furnham, & Graham, 1981; Jones, 1990). For example, a person taking part in one of our HRI studies is probably open to new technologies and not afraid of being in a user study situation in general. People also avoid situations either because they are not interested in the goals that can be achieved in the situation or because they fear that they cannot cope with the situation (Argyle, Furnham, & Graham, 1981). Therefore, as Jones (1990) reveals, it is difficult to maintain a strict dichotomy between choosing a situation and behaving in a situation.

Moreover, the situation itself is partially controlled or elicited by our own actions, i.e., our own actions in a certain situation trigger a certain response. Therefore, the interaction goal of the perceived person has to be taken into account which again is strongly bound to the situation. Often, recognizing the interaction goal is part of recognizing the situation from the perceived persons point of view.

Most interaction goals are at least in part concerned with the actors' attempt to manage the impressions that others form of them. For example, participants in HRI studies might be really upset because the robot does not perform as it should. However, they show their disappointment only to a limited degree since the experimenter is also around and people are usually polite in such situations. These self-representational goals are social interaction goals. Interaction goals, however, can also be task-oriented. Wrede, Kopp, Rohlfing, Lohse, and Muhl (to appear) have shown that the goal of the interaction determines the behavior of the actors in asymmetric interaction. Asymmetry in HRI refers to the different abilities of human and robot. While at some points of the interaction the goal of the robot is understanding exactly what the user intends to do, which can be achieved with clarification questions, at other points the goal might be to conceal misunderstanding to keep up the flow of the interaction. The decision between the strategies has to be based on the situation.

The influence of the situation has also been underlined from a developmental perspective:

> *"It is in actual situations that we meet the world, form our conceptions of it, and develop our specific kind of behavior for dealing with it. Situations present, at different levels of specification, the information that we handle, and they offer us the necessary feedback for building valid conceptions of the outer world as a basis for valid predictions about what will happen and what will be the outcome of our behaviors. […], behavior takes place in situations; it does not exist except in relation to certain situational conditions and cannot be understood and explained in isolation from them." (Magnusson, 1981b, p.9f.)*

Magnusson (1981b) shows from a developmental perspective that people learn in situations for other situations to come. This also highlights the importance of situational factors for the interpretation of action. The underlying process is called "ongoing, reciprocal person-situation interaction process" (Magnusson, 1981b, p.10). This process is the basis for understanding the behavior of a person in a given situation.

A common term to describe the strong relationship between person and situation is *situatedness*. According to Rohlfing, Rehm, and Goecke (2003) it "refers to specific situations in which actions take place" (p.133). A person or an agent "interacts with the situation by perceptual and effectorial processes" (Rohlfing, Rehm, & Goecke, 2003, p.136). Perceptual processes are all processes connected to sensory perception and its interpretation, whereas effectorial processes are cognitive and motoric processes controlling the agent's actuators. These processes are closely related to the context (see Section 2.1.2). According to this definition, agents are only interested in situational stimuli that they can perceive with their sensors. For example, the robot BIRON pays attention to visible stimuli but does not take into account the smell of the

environment. HRI differs from HHI in this respect which is just one aspect leading to asymmetry in the interaction.
From a robotics perspective, a robot is socially situated when it interacts with people (Dautenhahn, Ogden, & Quick, 2002). In this sense, situatedness only refers to action in a certain situation. Another view by Pfeifer and Scheier (1999) claims that

> *"An agent is said to be situated if it acquires information about its environment solely through its sensors in interaction with the environment. A situated agent interacts with the world on its own without an intervening human. It has the potential to acquire its own history, if equipped with appropriate learning mechanisms." (Pfeifer & Scheier, 1999, p.656)*

From this point of view, an agent can only be situated if it has a direct percept of the world that enables it to learn by acquiring an own history like humans do.
All definitions have in common that actors are situated because they act in a specific situation. For the following, situatedness means that actors always act in specific situations which they perceive with the sensors (senses) that they possess. Actors also learn from their actions, i.e., they memorize previous situations and can later exploit their memories and the knowledge connected to them. Hence, not only action but also learning is situated. It is assumed that this is also true for HRI where users learn about the interaction and the robot's abilities while they interact.
How appropriate the predictions of an actor are in a situation depends on the *situation awareness* (SA). Endsley has defined this concept as follows:

> *"The formal definition of SA is 'the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future'" (Endsley, 1988, p.97)*

Accordingly, Endsley differentiates three levels of situation awareness: perception (Level 1), comprehension (Level 2), and projection of future states (Level 3). Each of the levels has a specific degree of uncertainty and, related to it, a specific degree of confidence. The goal of the actor is usually to resolve uncertainties which, according to Endsley, Bolté, and Jones (2003), can be achieved by searching for more information, reliance on defaults, conflict resolution, thresholding (not reducing uncertainty past a certain point), bet-hedging, contingency planning, and narrowing options (reducing possible options for what may occur). Also, other actors can contribute to situation awareness by providing information and confirming one's own situation awareness. Despite the name, "situation awareness" is more closely connected to the concept of context than to the concept of situation because all three levels of SA involve cognitive processes which put them into context. Therefore, it will be further discussed in 2.1.2.
To sum up, the main aspect that characterizes perceived situations is that every agent understands situations differently based on how he or she construes them. The construction process is influenced by the information that is taken into consideration. This information

changes dynamically based on experiences that an agent makes and, hence, also the perception of the situation is dynamic and evolves even if two situations can be identified as being similar. Thus, by experiencing situations, agents learn. In the case that many people act the same in a situation even though each of them construes it in his or her own terms, it can be assumed that the situation and not the personality mainly influences the behavior. By their action, the people themselves change the situation in a way that contributes to their interaction goal.

2.1.1.3 HRI as a social situation

In this section it is shown how the situation concept translates to HRI and an approach to describe HRI situations is developed for further analysis. In previous work on this topic, Fischer (2000) has analyzed experimental data in terms of how the subjects conceptualized an HRI situation. She has reported that beliefs about the communication partner showed in linguistic behavior. Moreover, the situation influenced the language the actors used.

> *„the speakers have been found to constantly define and, if necessary, redefine what they understand the situation to consist in, depending on their current hypotheses about their communication partners and on their own emotional state.“ (Fischer, 2000, p.7)*

Thus, Fischer (2000) again points out the person-situation interaction process and its existence in HRI. Users seem to form hypotheses of their robotic interaction partner and attribute occurrences to the robot as an entity depending on the situation. Hence, the robot seems to be perceived as a social actor acting in a social situation. Goffman (1961) defines the term *social situation* as follows:

> *“By the term social situation I shall refer to the full spatial environment anywhere within which an entering person becomes a member of the gathering that is (or does then become) present. Situations begin when mutual monitoring occurs and lapse when the next to last person has left” (Goffman, 1961, p.144)*

According to Goffman’s (1961) view, a social situation exists when at least two people share the same space and monitor each other, i.e., they interact. This is also true for HRI situations, even though the actors are not two people but a person and the robot. HRI also often follows the same episode structure as social encounters in HHI (Argyle, Furnham, & Graham, 1981) as the following examples from the home tour scenario show (in parenthesis: utterances of the users):

1. greeting (“Hello Biron”)
2. establishing the relationship, clarifying roles (“I will show you some objects today”)
3. the task (“This is a cup”)
4. re-establishing the relationship (“Now let’s turn to the next task”)
5. parting (“Good bye, Biron.”)

Wagner and Arkin (2008) define social situations in HRI as "the environmental factors, outside of the individuals themselves, which influence interactive behavior" (p.278). Furthermore, the authors (2008) "currently know of no direct consideration of the theoretical aspects of social situations as applied to interactive robots", even though most researchers focus on certain situations that occur within the limits of the scenarios. Wagner and Arkin (2008) introduce interdependence theory that represents social situations computationally as an outcome matrix. In the context of a human-robot clean-up situation, they vary the number of victims and hazards in order to calculate the characteristics of the HRI situation and to predict reward and cost outcomes. However, they do not identify general characteristics of situations. In social psychology several methods for analyzing such characteristics in social situations exist (Argyle, Furnham, & Graham, 1981). Among others these are the dimensional, componential, and the environmental approach and the study of particular kinds of social behavior. The *dimensional approach* assumes that situations can be described or compared along certain dimensions, whereas the *componential approach* is based on the belief that social situations are discrete, not continuous entities. This approach seeks to determine the components in each situation type and to understand their relationship. It further assumes that some elements are common to all situations but they take different forms. Hence, situations can be compared at the level of single elements. However, this approach does not result in a comprehensive list of situational elements to explain their interaction. The main concern of the approach is to describe how various components of situations vary across situations. The *environmental approach* concentrates mainly on the physical aspects of situations and seeks to explain how certain elements of the environment are perceived or related to behavior. Finally, the *study of particular kinds of social behavior* is a more limited approach. It exclusively seeks to investigate the influence of salient situational variables on a particular kind of behavior.

Argyle, Furnham, and Graham (1981) propose a functional view of situation: people enter situations to achieve goals that satisfy their needs. The authors introduce nine features for the analysis of social situations that will be explained in the following. These features can be viewed as a mainly componential approach which the author finds useful to describe situations and to determine differences between them. In the following, the components are introduced and, thereafter, they are related to HRI.

Goals and goal structure

As mentioned before, Argyle, Furnham, and Graham (1981) view situations as occasions for attaining goals. Most social behavior is goal directed. Knowing the goals of the interactors provides the basis for understanding their behavior. A person can pursue more than one goal at a time. The interrelation of the goals is then called goal structure.

Rules

"Rules are shared beliefs which dictate which behaviour is permitted, not permitted or required" (Argyle, Furnham, & Graham, 1981, p.7). Rules regulate behavior by determining what is appropriate so that the goals can be attained. Some rules emerge during the life of a particular group; others are imposed on the members of a group from above. The rules differ

with groups and situations and can be broken. Argyle, Furnham, and Graham (1981) describe three main kinds of functions of rules: (a) universal functions that apply to all situations, (b) universal features of verbal communication, and (c) situation-specific rules, based on the goal structure of the situation.

Roles

With regard to expectations, roles are defined by the rules that apply to a person in a certain position. They provide the person with a fairly clear model for interaction and require special skills. Roles can change within situations and they are interlocked with certain goals, rules, the repertoire of elements, physical environments, and certain difficulties.

Repertoire of elements

Elements provide the steps needed to attain goals. The repertoires of elements are restricted and vary in different situations. No one set of behaviors exists that covers all situations in detail. Typical sets of elements for certain situations can be defined and are usually more useful. However, it is necessary to decide on the units to be coded (for example, complete utterances or parts of them, facial expressions or interpersonal distance in certain time intervals, etc.). In general, verbal categories and bodily actions vary strongly between situations, whereas non-verbal elements are used in every situation to a different degree (Argyle, Furnham, & Graham, 1981).

Sequences of behavior

The elements in a situation may have a certain sequence. Some very common sequences are adjacency pairs, social-skill sequences, and repeated cycles of interaction. Also, the episodic structure of situations is sequential, i.e., the main task consists of subtasks with a special order. Transition probabilities between actions can be identified using ethological methods (for example, A has a high probability of leading to B, C has a high probability of leading to A). However, this approach has not been very successful in the analysis of human social behavior because human sequences are too long and too complex and earlier parts of the sequence can influence later parts. Moreover, sequences strongly depend on the situation.

Concepts

According to Argyle, Furnham, and Graham (1981), interactors need shared concepts for handling situations. In social situations, the interactor needs categories in order to classify persons, social structure (role, relationship between people), elements of interaction (for example, classification into friendly/hostile), and relevant objects of attention (parts of the physical environment, task-related objects). The actors need to interpret the behavior of the other interactors and to plan their own behavior.

Environmental setting

Social situations have *boundaries* of action, i.e., physical enclosures in which behavior takes place, such as rooms. All boundaries contain *props* (for example, furniture, objects). Each prop

has a particular social function and often a special meaning attached to it. *Modifiers* are physical aspects of the environment or, in other words, the qualities and quantities of conditioners within the boundaries (for example, color, noise, and odor). *Spaces* refer to the distances between people or objects which usually convey a certain meaning (for example, being close to another person indicates liking). The environmental setting determines which behaviors are appropriate within it. Moreover, the settings substantially influence the perception of the situation and people's actions.

Language and speech
As Fischer (2000) found for HRI, how people construe the situation influences their linguistic behavior. Also Argyle, Furnham, and Graham (1981) identified language and speech as one factor for the characterization of social situations. They divide between language, which is the underlying system of grammar, and speech, which is the way people actually talk. Generally, linguistic features are associated with every situation. Some situations constrain language much more than others, for example, because of abilities of the opponent or certain rules.

Difficulties and skills
Some situations are difficult because of the stress that they cause (for example, job interview, experiment situation). Difficulties in social situations are a function of the skills of a person. If a person is skilled, the difficulties are smaller.

These components show that all situations vary along the factors. In the following, different levels of specificity are proposed that can be used to describe an HRI situation and relate the factors to the model which is depicted in Figure 2-1. The levels of specificity that will be differentiated here are:

0. characteristics of situations in general (lowest level of specificity),
1. typical characteristics of HRI situations,
2. scenario-specific characteristics,
3. task-specific characteristics,
4. and characteristics of every specific situation (highest level of specificity).

Characteristics of situations in general are that they have a certain timeframe (a beginning and an end), they are dynamic, and they take place one after another. All situations exist physically in the world around us. Also actors are part of the physical situation. However, each actor perceives the situation individually, based on former experience, personality, and interaction goals. These general characteristics apply to all kinds of situations and also to HRI. HRI situations are one specific kind of situation (just as HHI is another one) with specific characteristics.
Characteristics of HRI situations primarily concern characteristics of the actors – the human and the robot. Similar to HHI, both the robot as well as the human are embodied and share the same physical space. For the situations that contribute to the model, it is further assumed that

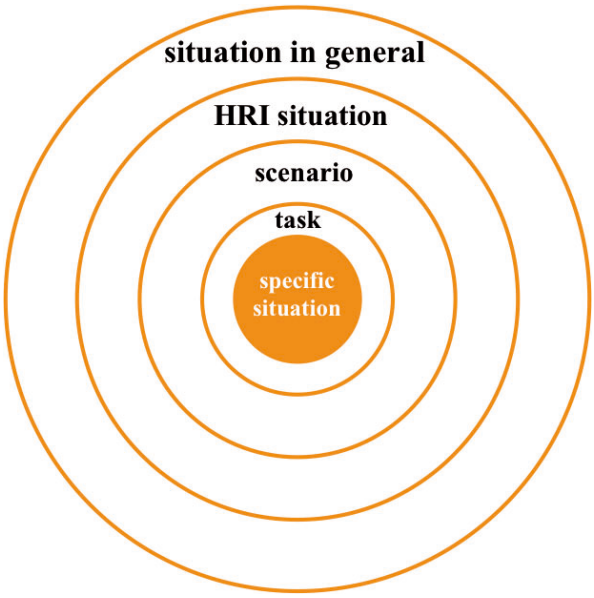


Figure 2-1. Levels of specificity of the situation
The degree of specificity increases from the outer to the inner circle.

both actors act autonomously. However, as has been reported in Wrede et al. (to appear), HRI is asymmetric because the robot's and the user's modalities and abilities differ. For example, robots without arms cannot use gestures, some robots are not mobile, and most robots cannot perceive smells. Because robots have different modalities and "sensors" to humans, they perceive the situation differently. Apart from the modalities, also the mental abilities vary. Even though a robot might speak with the user, most robots only have a very basic vocabulary and need more time for computation than humans do. The repertoire of elements of the robot in general is restricted to a certain set that, usually, is smaller than the set of a human and also timing is often asymmetric in HRI situations. For this reason, turn-taking in HRI might differ from HHI. The language and speech factor in this case is restricted by the abilities of the robot that forms a part of the social situation as such.

Another characteristic of HRI situations concerns the fact that even though robots are social interaction partners, they are also machines. This is important to keep in mind because humans have different attitudes towards machines than towards other humans. People who are afraid of using a computer (or a robot) are certainly not necessarily afraid of talking to another human. On the other hand, some people have social relationships with or via their computers but avoid relationships with people in real life. People's attitudes toward HRI might also be connected to their experiences. To date, only a few people have experienced HRI situations with social robots. That is why they do not have any concrete knowledge about them but need to transfer to HRI what they know from HHI and interaction with other machines.

The next level of specificity depicted in Figure 2-1 is ***scenarios***. As described in Section 1.2, the scenario that serves as an example here is the home tour. The scenario in HRI determines what roles the actors take. The home tour focuses on the teaching of the robot. The general goal is that the robot acquires knowledge about objects and rooms. Hence, the robot is the learner while the user is the teacher. These roles imply rules for the interaction, for example, that the user is free to guide the robot around in order to teach it some object or room and the robot follows.

This also requires that the users are more skilled with respect to the subject, which is the home and objects within it, and that they know how to teach the robot. The scenario also implies that the robot has the skill to learn (or at least to give the feedback that it has learned something) and to drive around. Hence, scenarios are connected to certain abilities of the robot. They are also intertwined with the environmental setting. The boundary of the setting is the apartment. It contains props such as furniture and the objects the robot needs to learn, and modifiers, for example, the color of the walls. These factors with respect to the scenario describe given aspects of the situation.

From an even more specific point of view, situations can be described with respect to the **tasks**. On the task level, also the goals and the goal structure are more specific. One goal of the human might be to show a certain object to the robot. Showing the object might include subtasks such as getting the robot's attention, saying the name of the object, and receiving an appropriate feedback. On this level of specificity, concrete repertoires of elements and sequences of behavior can be identified which will be demonstrated in Chapters 4 and 5. Also typical difficulties can be recognized which will be achieved here with Systemic Interaction Analysis (Sections 3.2.6, 5.2, and 5.3). On the task level, language and speech are even more restricted since the robot only does certain things as a reaction to specific commands, for example, the amount of commands to make the robot follow the user is limited.

On the level of a ***specific situation*** all factors describe a concrete part of the interaction with a certain beginning and end. Every situation can be described on this level. The description includes concrete actions, specific elements chosen from the repertoire in a certain sequence, specific goals, a specific environment, and specific difficulties.

Finally, it needs to be considered that most situations researched in HRI are not only HRI situations but also *study or experiment situations*. Therefore, the humans are not only users of the robot but also participants in the study. In this role, they will probably want to do a good job or to fulfill the expectations of the experimenter. If the experimenter is present, the social behavior of the subjects might change. That is the reason why the experimenter in all the studies presented here remained in the background (see Section 1.4). It was not always possible that he or she would leave the room due to security reasons; actually, the feeling of being alone in a first-contact situation in a foreign environment with a mobile robot might also scare the participants which might influence the interaction more than the presence of the experimenter. Finally, the study situation could also be stressful because people are in the *focus of attention* and they might be afraid of *failure.* According to Argyle, Furnham, and Graham (1981) both these factors cause difficulties. This was considered when designing the user studies in order to keep difficulties to a minimum.

To conclude, it has been shown that HRI can be analyzed as a social situation on different levels of specificity. The lower levels of specificity (situation in general, HRI situation) allow for a general comparison of HRI with other types of interaction (for example, HHI and interaction with virtual agents). The scenario level enables comparison between HRI scenarios, the goals that they include, the roles and the rules that they imply, the skills that are necessary, and the environment that they should be situated in. The task level allows for the comparison of

behavior within and between tasks. Behaviors that are used within the tasks can be identified and compared to behaviors that occur in the context of other tasks. The highest level of specificity, specific situations, allows for in-depth descriptions of situations and for the identification of similarities between situations.
In the following, the levels of situation in general and HRI situations play a role insofar as at various points parallels to HHI are drawn to show relations and distinctions. The scenario level allows to compare different scenarios in HRI such as the home tour (see Section 1.2). The task level and the level of specific situations are the main levels of analysis here, because the goal is to find out what the users do in specific situations, how they try to complete tasks, whether their behavior follows certain sequences, and which difficulties arise.

2.1.2 The concept of context in HRI

At the beginning of this chapter, it was noted that the terms context and situation are often used synonymously. The context was shortly defined as the internal knowledge of the actors that is needed to handle situations. In the following, the concept of context will be further distinguished to the concept of situation and introduced to HRI.
What authors refer to as context often remains unclear (Butterworth, 1992), even though there seem to be different usages of the term. Cole and Cole (1989) suggest a *cultural-context view*. From this viewpoint, cultures provide guides to action by means of their language and their material structure. Moreover, cultures arrange the occurrence of specific contexts. The frequency with which contexts are encountered influence skills and schemes that people develop. For example, most people in Germany are trained to shake hands when greeting a business partner, while people in Japan bow. In general, people are skilled in their own cultural habits. However, cultures are not necessarily completely separated as individuals may participate in several cultural contexts simultaneously. Cole and Cole's (1989) view on cultural context also includes the assumption that the context in which a person is provided with a task is influential. As stated before, the culture arranges the occurrence of these contexts. Hence, in some cultures a certain mathematical problem might arise primarily in the classroom and the children might be able to solve it there. In contrast, in other cultures the same problem might come up when trading goods and someone might be able to solve it in this context but not in the classroom where it is much more abstract.
The cultural context view is just one approach to the concept. Focusing at *task analysis*, context is seen as a source of information about how people carry out tasks. It describes available information and opportunities for action, the frequency of information presentation, and information about factors that have or do not have to be taken into consideration when making decisions (Shepherd, 2001). In this sense, context describes all knowledge that supports task solving. Since tasks are in the center of the following analyses, this definition strongly influences how context is understood here.
The term context is also used in linguistics where it refers to specific parts of an utterance near a unit that is in the focus of attention (Mercer, 1992). According to Gumperz (1982), utterances can be understood in many different ways because people make decisions on how to understand them based on their understanding of what is happening in the interaction at a certain point of

time. In the interaction process, partners implicitly communicate what Gumperz calls *contextualization cues* which are linguistic features that signal contextual presuppositions. Some examples of contextualization cues are dialect, style switching, and prosodic phenomena. Following this linguistic point of view, Gumperz defines contextualization as

> *"the process by which we evaluate message meaning and sequencing patterns in relation to aspects of the surface structure of the message, called 'contextualization cues'. The linguistic basis for this matching procedure resides in 'co-occurrence expectations', which are learned in the course of previous interactive experience and form part of our habitual and instinctive linguistic knowledge." (Gumperz, 1982, p.162)*

Accordingly, context cues from a linguistic point of view are not *what* we say but *how* we say it. They have been learned in previous interaction situations and are applied instinctively. Since we have lots of communicative experience, we also have a huge body of contextual knowledge. What part of this background knowledge is relevant changes during the interaction (Gumperz, 1982). However, the following analysis focuses on what the users say and neglects how it is said. Therefore, the linguistic point of view is not appropriate here.

In contrast, the definition by Rohlfing, Rehm, and Goecke (2003) influences the definition of context that will be used. The authors describe context as a general construct (more general than situations) that can be defined on at least two levels. These two levels are the socio-cultural (global) context and the local context (for example, the context of an experiment situation). Especially the local context is of importance for the following analysis. Furthermore, the authors divide context into intracontext and intercontext. Intercontext is a complex network of interacting subsystems. The action as such is part of the intercontext. This is what will be analyzed in the following because it can be observed. In contrast, the intracontext is established by the agent itself. It allows the agent to make sense of activities. Establishing an intracontext is equivalent to learning. The more sophisticated the intracontext, the more sophisticated actions can be conducted by the agent. Intra- and intercontext influence each other.

To conclude from all the positions presented above, for the following analyses the term context refers to local, dynamic repositories of knowledge gained in former situations that help to understand new situations. As mentioned above, this definition is mainly based on the view from task analysis. This usage of the term implies that context is learned from situations, it is based on situations, surrounds all situations, and evolves with situations. When a person perceives a situation, various contexts are taken into consideration. In other words, every specific situation is embedded in a number of different contexts that may overlap. Figure 2-2 depicts possible relations between the situation and the contexts. Context 'a' barely overlaps with the situation and the other contexts. Context 'b' has a much bigger meaning for the situation and is more closely connected to context 'c' that surrounds the situation completely. Finally, context 'd' is also barely connected to the situation but it overlaps with context 'c'. Consequently, contexts 'b' and 'c' are most certainly taken into account in the situation, while the consideration of contexts 'a' and 'd' is less probable.

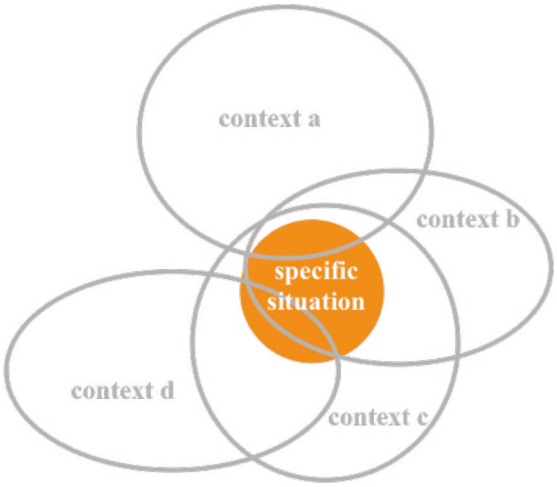


Figure 2-2. Relation between specific situations and contexts

What contexts are considered depends on the perception of the situation and memories of other situations that this new experience seems similar to (Schank, 1999). In turn, the contexts determine what comes to the mind of an interlocutor in a certain situation; for example, certainly different robots come to people's minds when talking about movies, Bielefeld University, a concrete HRI study, or cleaning the home. Thus, the contexts drive reminding. According to Schank (1999), reminding occurs in context and across context.

> *"When it occurs in context, we learn more detail about the context itself and we enlarge and store our knowledge base by making generalizations and later recalling expectations to those generalizations. When reminding occurs across context, we are able to use what we have observed in one domain to help us in another. Such leaps are the stuff of intelligence and insight." (Schank, 1999, p.88)*

Accordingly, single contexts are generalized, knowledge from one context is used in other contexts when the person is reminded of it and, consequently, people learn. What interactors are reminded of strongly depends on the situation and their experiences. People have so-called *prototypes* of most events because they have experienced a certain event many times and have generalized over all occurrences (Schank, 1999). Hence, people are often reminded of prototypes and not of specific situations. What prototype is triggered depends on the classes of reminding (Schank, 1999, p.22) which are:

1. physical objects can remind of other physical objects
2. physical objects can remind of events
3. events can remind of physical objects
4. events can remind of other events in the same domain
5. events can remind of other events in different domains
6. new experiences can remind of prototypical cases
7. abstract ideas can remind of related ideas

The overview shows that there are many sources of reminding that put people's actions in context. It again points to the dynamic view of remindings, because they are construed depending on the sources that come into play.
In HRI research, context has been investigated in connection to situation awareness (SA) because the considered contexts influence SA. Fischer (2000) has researched the SA of the user. Fischer and Lohse (2007) investigated what naïve users think about the SA of the robot. They conclude that humans estimate a robot's SA based on their mental model of the system. Their beliefs about the robot's SA can be shaped with the help of verbal output. This work points out that not only the own SA but also the SA of the interaction partner is of interest, especially if the interaction partner is an agent with capabilities that differ from human capabilities. To understand the system, the human has to determine what context the system does take into account. Most probably the user assumes to a certain extent that the system has similar depositories of knowledge gained in former interaction situations, because this shared intercontext would improve the interaction. The fact that the robot has different abilities than the user but can shape the users' expectations is one reason why HRI can contribute to expectation research as has been specified in the fourth aim of this thesis. In this regard, the special advantage of HRI is that the robot can repeat the same behaviors in the interaction with different users. Thus, it can be determined how they are interpreted and what contexts each user takes into account. In general, the following contexts can be inferred for HRI situations:

- general interaction context (includes knowledge about interaction [in a certain culture] in general, i.e., how to behave in interactions)
- agent contexts (knowledge about the robot, the user, [the experimenter]).
- task contexts (for example, context of greeting somebody you meet for the first time, knowledge about teaching tasks)
- goal contexts (for example, solve tasks, self-presentation, learn about robotics, have a good experience)
- spatial contexts (for example, knowledge about the apartment such as which room is the living room)
- object contexts (for example, what is a table, how to show a table [which should usually not be lifted up like a cup when it is shown])

Since the studies presented below are first-contact situations, the users do not have experiences from HRI. That is why they depend on contexts from HHI and other situations. Each of the contexts can serve to support other contexts. For example, if the user does not have much context with respect to the robot as an agent, the agent context with respect to human interaction partners might still help to understand what the robot is doing, especially since the interaction with the robot BIRON is designed based on HHI.
However, which contexts are applied at a certain point of time is hard – if not impossible – to infer when looking at the video data. Nevertheless, based on the contexts, people perceive the situations in a certain way and act according to how they understand what is happing. Therefore,

how users perceive the interaction situation can be inferred from their actions and from the questionnaire data (see Chapters 4 and 5).

2.2 Expectations and expectancies

In the last sections, it has been argued that situations remind us of other situations which are represented in certain contexts. We have *expectations* of what will happen because it has happened before in the prototypical situations that make up the context. Expectations have played an important role in research (Blanck, 1993). In the literature they have also often been called *expectancies* (Feather, 1982c) and both terms are used in many ways in everyday life: employees have expectations toward their employers, the society has expectations toward politicians, someone can expect a baby, and measure life and health expectancies. The effects of expectancies were examined in contexts like teacher/student relationships, influence of court-room judges on juries' expectancies, relation between expectancies and racial inequity, and usage of expectancy effects in organizations to enhance work motivation.

In the following, the concepts "expectation" and "expectancy" are introduced and, based on expectation theory, it is explained why expectations play a crucial role in HRI. It is further described how expectations emerge, what their functions are, how they influence information processing, and how they relate to other concepts, i.e., beliefs, scripts, schemas, and attitudes. Throughout the section, examples are provided that show the relation between expectations and HRI.

The title of this section is "Expectations and expectancies". Before taking a closer look at the theory, it shall be clarified why in the following both terms are considered. Feather (1982c) found that in the literature both termini are applied interchangeably. While he does not point out differences between the two terms, LePan (1989) differentiates them with regard to specificity. He provides the following example:

> expectancy: "I will be rich one day!"
> expectation: "My influence on this person will assure that I get the contract which will make me rich."

While the expectancy is abstract and general, the expectation is much more specific; or in LePan's (1989) words:

> *"A sense of expectancy is unlike an expectation [...] in that it bypasses any chain of circumstances and events, and deals only with end results: it looks forward not to a specific event, but to a final condition." (LePan, 1989, p.75)*

This means that the expectation is the answer to the specific question "what will happen next?". In contrast, the expectancy is related to the more general question "will it happen?". Jones (1990) supports the view that "expectancy" is the more general term:

> *"I would like to collect all manner of expectations, predictions, hypotheses, and hunches about a target person's upcoming behavior under the commodious label 'expectancy'" (Jones, 1990, p.78)*

From this point of view, expectancies are not only more general than expectations but actually contain expectations as one concept among others such as predictions and hypotheses about other people's behavior.
The following analysis focuses on the question *how* users try to reach a certain goal. To answer this question, a higher degree of specificity is necessary. Therefore, the term "expectation" will be preferred. However, when referring to previous work or when theoretic work clearly focuses on the outcome of an event, the term expectancy is used.

2.2.1 Definition and characteristics of expectations

In a recent paper, Roese and Sherman (2007) define expectancies as follows:

> *"Expectancies are beliefs about a future state of affairs, subjective estimates of the likelihood of future events ranging from merely possible to virtually certain. […] The expectancy is where past and future meet to drive present behavior" (Roese & Sherman, 2007, p.91f.)*

This definition indicates that past experiences shape expectancies which, in turn, support anticipating the future and shape the behavior of a person. Olson, Roese, and Zanna (1996) call the kind of expectancies defined here *probabilistic expectancies* since they describe what could happen with a certain probability ranging from being possible to being certain:

> *"[…],* ***probabilistic expectancies*** *refer to beliefs about the future, or perceived contingency likelihoods (what might happen). […],* ***normative expectancies*** *refer to obligations or prescriptions that individuals perceive for themselves or others (what should happen)." (Olson, Roese, & Zanna, 1996, p.212)*

Probabilistic expectancies are hypotheses about what might happen. If the probability that something happens is 100%, the hypothesis is *factual*; if it is lower than 100% the hypothesis is *subjective*. In contrast to probabilistic expectancies, normative expectancies are not focused on future events. Rather they describe what actors should do in a certain situation based on the norms of a certain culture and on the cultural context (see Section 2.1.2).
Both forms of expectations are important for the purpose of analysis. Of primary interest are the questions what the users expect the robot to do in response to their actions and what they do to obtain a certain reaction. These questions concern probabilistic expectations. An example from HRI shall show this connection. When users greet the robot, it can be assumed that they would, for example, say "hello", and expect the robot to greet them, too. Meeting the robot in their home country, they might, among others, have the probabilistic expectation that the robot speaks their language. The expectation in this case can be inferred from HHI and the interaction

context (see Section 2.1.2). Also normative expectations play a role in this case, for example, in Europe the robot should answer the greeting of a user by saying something polite like "hello" and not "What's up?" as it might be appropriate among friends, and it should not bow. Depending on the situation, it should possibly shake hands. Burgoon (1993) has defined these specific as well as general expectancies from a communication science point of view:

> *"'Expectancy' in the communication sense denotes an enduring pattern of anticipated behavior. These expectancies may be general – pertaining to all members of a given language community or subgroup – or particularized – pertaining to a specific individual. In the former case, they are grounded in societal norms for what is typical and appropriate behavior. In the latter case, they incorporate knowledge of an individual actor's unique interaction style, which may differ from the social norms." (Burgoon, 1993, p.31)*

Accordingly, a particular individual and culture impact the expectations of an interaction partner. Both influence normative expectations. While the content of the normative expectations varies depending on the culture, their structure should be pancultural (Burgoon, 1993, p.32). However, Burgoon stresses that individuals do not necessarily have to comply with societal norms (see Section 2.2.4).

Both, normative and probabilistic expectancies can be described with certain properties. Olson, Roese, and Zanna (1996) discuss four properties: certainty, accessibility, explicitness, and importance. In their recent work, Roese and Sherman (2007) examined five slightly modified parameters that reflect new theoretical insights in the field: likelihood, confidence, abstractness, accessibility, and explicitness. In the following, these parameters are elaborated.

Likelihood of occurrence is a basic way of describing expectancies by their degree of probability. Likelihood is determined by input sources such as past experience, social learning, and the media. Mood also influences likelihood. Positive mood increases the perceived likeliness of positive events, and negative mood increases the perceived likelihood of negative events. The subjective likelihood of expectancies can increase if they are confirmed. With reference to the example of users greeting a robot, the subjective likelihood of the robot greeting the users when they say "hello" will increase if the robot has appropriately answered the users' greeting before. In contrast, it will decrease if the robot has not greeted. It might decrease even more if the users become frustrated.

The parameter *confidence* is orthogonal to likelihood. High confidence is not the same as high likelihood. "Both low- and high-probability events may be expected with both high or low confidence" (Roese & Sherman, 2007, p.94).

The next parameter discussed by Roese and Sherman (2007) is *abstractness*. It refers to the difference between concrete and specific representations and abstract generalizations that summarize experiences from several events, people, and contexts over time. Research has shown that events that are imminent are conceptualized more concretely than events that are in distant future, because people value events more if they are temporally closer. *Accessibility* determines how easily an expectancy is brought to conscious attention and how likely it is to

influence judgment. Accessibility increases if an expectancy is frequently activated or has recently been activated in memory. Expectancies that are disconfirmed are more accessible since sense-making activities are triggered (see Section 2.2.3). It shall again be referred to the example of a user greeting a robot. As mentioned before, greeting is a very natural action in HHI. The expectation that someone replies to a greeting is deeply anchored in a greeting script (for further elaboration on the term "script" see Section 2.2.5.3). The script itself is often accessed in every-day situations. Therefore, the expectancy should be highly accessible also in HRI. If the robot does not reply to the greeting, the expectancy is disconfirmed and the user should involve in sense-making activities, i.e., try to find a reason why the robot did not respond. Disconfirmation and sense-making processes will be further discussed in Section 2.2.3.
Finally, expectations have a certain degree of *explicitness*. Explicit expectancies can be consciously reported while implicit expectancies are held unconsciously. Most expectancies are held without the person being aware of them. They only become explicit when the person has to articulate them. However, the verbalization might not exactly represent the expectation itself, because it is influenced by other factors such as the promotion of the own personality, or the person might simply not be able to identify the expectations that led to a certain behavior. This issue is called *unwilling and unable problem* (Eagly & Chaiken, 1993). It motivates the question of how to measure expectancies which will be discussed with respect to HRI and the home tour scenario in Chapter 3.
To conclude, the parameters support the description of expectations in the first-contact HRI situations analyzed below. Since likelihood depends on (dis-) confirmation it should increase or decrease during the interaction with the robot if it confirms or disconfirms the users' expectations. Whether this connection can be found in the data will be analyzed in the following. During the interaction, more generalizations across situations should occur and the expectations should shift from being concrete to more abstract. The accessibility of expectations from HHI in HRI is higher if they are activated frequently. It can be expected that expectations which are activated a lot in HHI, such as expectations connected to greeting, are very accessible. Finally, whether the expectations are explicit and the users are able to report them will be determined with the help of interviews (see Section 5.4).

2.2.2 Formation of expectations

Some of the examples have already touched on the question of formation of expectations. In the following, this issue will be discussed in detail. One classic experiment in social psychology that illustrates the formation of expectations was reported by Kelley (1950). In the experiment, students were told that a guest instructor was either warm or cold hearted. After the lecture, students judged the instructor. Their judgments clearly showed assimilation effects: the group of students that were told that the instructor was warm hearted, rated him as being more considerate, informal, sociable, popular, good-natured, humorous, and humane. These students were also more open for a discussion with the instructor (56% in the warm hearted condition vs. 32% in the cold hearted condition). Another classic example for this phenomenon is Langer and Abelson's study (1974) on therapists' impressions of "patients" and "job applicants" (see Biernat (2005) for an overview of work in this field).

In these examples, the expectations were primed by *communication from other people* before the actual encounter (see also Section 2.2.5.2 on priming of schemas). Besides communication from other people (indirect experience), Olson, Roese and Zanna (1996) name two more sources of expectations: *direct personal experience* and *beliefs that are inferred from other beliefs*. Every expectation is based on at least one of these sources, all of which can be biased (Darley & Fazio, 1980). Direct personal experience tends to have a greater influence on expectation formation than indirect experience (Fazio & Zanna, 1981) and leads to semantic and episodic expectancies. According to Roese and Sherman (2007), *semantic expectancies* are: "preexisting knowledge structures that are extracted from ongoing experience, stored in memory, and retrieved when needed." (p.95). In contrast, *episodic expectancies* are based on single events that someone experiences. While semantic expectancies are an efficient way of using past experiences, episodic memories provide depth by including information from the concrete situation. Semantic and episodic expectations interact and can be applied in parallel. Usually, people at the outset have specific expectations, which become more general over time, and are finally balanced at mid-levels. Based on this concept, one can assume that expectations change over time depending on the situation. Since users in first-contact situations do not have any experience in interacting with the robot, their confidence, for example, that it answers when greeted, should be low in the beginning but increase during the interaction if the expectation is confirmed. Users should first form episodic memories of the robot's behavior. For the analysis of the studies below, it is thus hypothesized that users are able to reflect concrete experiences in much more detail after a first-contact situation with a robot based on their episodic memories than after long-term interaction.

It can be concluded that the situation influences the expectations. Heckhausen (1977) has introduced a model that depicts this connection. The model encompasses a four-stage sequence of events (see Figure 2-3). Based on these stages, Heckhausen (1977) names different kinds of expectancies, each of which can be allocated to one stage of the model. From top to bottom the respective expectancies are: *situation-outcome expectancies, action-outcome expectancies* (these two are typically addressed in motivation theory, for example, by Tolman, 1932), *action-by-situation-outcome expectancies*, and *outcome-consequence expectancies*.

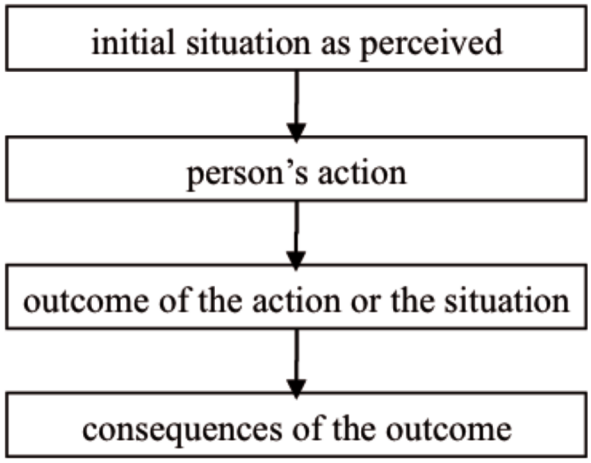


Figure 2-3. Model of the influence of the situation on expectancies (Heckhausen, 1977)

Out of these four kinds of expectancies, Maddux (1999) focuses on action-outcome expectancies and situation-outcome expectancies which he calls *behavior-outcome expectancy* and *stimulus-outcome expectancy*, respectively. Behavior-outcome expectancies are beliefs that a specific behavior results in a specific outcome or set of outcomes. Behavior-outcome expectancies have been included, for example, in the theory of reasoned action (Fishbein & Ajzen, 1975) and the theory of planned behavior (Ajzen, 1988).[9] They can be distinguished based on the outcomes that can be external (environmental events) or internal (nonvolitional responses). In the case of greeting the robot, the user's action of saying "hello" is the behavior and the expected outcome is the robot's answer, which is an environmental event. Behavior-outcome expectancies are important with respect to the central questions of this thesis: what do users do and what do they expect the robot to do in response?

While behavior-outcome expectancies focus on behaviors that lead to certain outcomes, stimulus-outcome expectancies "are concerned with beliefs that certain events provide a signal or cue for the possible occurrence of other events" (Maddux, 1999, p.23). Thus, a stimulus signals whether a certain event will occur or not if a person engages in a certain behavior. Stimulus-outcome expectancies trigger behavior-outcome expectancies. Comparable to behavior-outcome expectancies, they can signal whether an environmental event (stimulus-stimulus expectancy) or a nonvolitional response (stimulus-response expectancy) will occur. It is interesting to note that the distinction between behavior-outcome expectancies and stimulus-outcome expectancies again points to the sub-units of a situation – event and stimulus – as introduced by Magnusson (1981a) (see Section 2.1.1.1). Also the robot can produce events, i.e. behaviors, and stimuli that signal the occurrence of other events to the user. Therefore, both are important for the following analysis.

A third kind of expectancy is the *self-efficacy expectancy* (Maddux, 1999). It describes someone's belief in the ability "to perform a specific behavior or set of behaviors under specific conditions" (p.24). The self-efficacy expectancy addresses cognitive abilities as well as motor abilities. In the example provided above, the self-efficacy expectancy of the users is that they are able to trigger an appropriate answer from the robot when greeting it. Since the robot interacts using natural modalities, i.e., speech as used in HHI, this expectancy should be rather high. However, it depends on the users' expectations towards robots in general and BIRON in particular. Accordingly, Jones (1990) distinguishes target-based and category-based expectations. If an actor forms an expectation during the interaction with a certain robot, it is *target-based*. An expectation toward BIRON could be that it understands speech because the users have experienced the robot understanding an utterance before. In contrast, prior expectations brought into the interaction are rather *category-based* which means that people have certain expectations because this is what they would expect of all members of that group. The robot BIRON belongs to a group of robots in general or social robots in particular. A category-based expectation towards the robot would also be that it understands speech *but* not

[9]Young et al. (2009) state that both theories can help to determine the willingness of people to adopt robots for application in domestic environments and, thus, stress the value of transferring theories from social psychology to HRI. The theory of reasoned action points to the utility, effectiveness and price of the robots. The theory of planned behavior points to the importance of perceived behavioral control; however, not only to control of the robot (which is especially important because the robot is physically present and might be mobile) but, for example, also to control of how owning it affects social status.

because the users have heard the robot speak before but because they assume that all robots (or social robots) do understand speech. In general, in HHI category-based expectations are discarded in favor of more target-based expectations during the interaction (Jones, 1990). Category-based expectations can be divided into dispositional and normative expectations (Jones, 1990). Dispositional expectations are based on the belief that different members of a group share certain characteristics (dispositions); for example, people participating in HRI studies are interested in robotics and technology in general, or social robots are able to understand speech. Dispositional expectations are more probabilistic than target-based expectations. They tell us what might happen but when the probability is not close to certain, allow us to turn to another expectation quickly. Dispositional expectations are a combination of several expectations about the situation while normative expectations are more strongly bound to the situation in focus.
Target-based expectations vary on a dimension from replicative to structural:

replicative: we expect a person to behave in the same way again if a situation is rather similar

structural: expectations based on a complex cognitive schema or an implicit theory of personality that allows predictions for people with certain traits

To conclude, expectation formation is guided by many factors such as the information that is taken into account. Once expectations are formed they are not necessarily stable but part of the dynamic memory that changes with the experiences that agents make. This is true especially for target-based expectations which are less probabilistic than category-based expectations and more likely to be replaced in the case of disconfirmation. For the analyses presented below, it seems reasonable to assume that the users' expectations are mainly target-based because they have not experienced the interaction with other members of the group, i.e., they have never interacted with a robot before. Therefore, in the model that is introduced below, expectations are regarded as being highly dynamic. Moreover, the expectations are assumed to be replicative, i.e., the users believe that the robot behaves the same in a similar situation because it has done so before. Thus, it can be assumed that the users do not change their behavior if it has turned out to lead to a satisfying result before. The users might also attribute traits to a robot which would be in favor of the structural dimension. However, they barely know the robot. Moreover, transferring different personality traits from HHI to HRI and applying complex cognitive schema would cause additional cognitive load. This is underlined by the finding that most participants found it quite difficult to judge the personality of the robot in a personality questionnaire.[10]

2.2.3 Function and processing of expectations

The goal of this section is to show that the function of expectations is closely related to the question of how they are processed in the human brain. Moreover, it describes the influence of expectations on the storage and retrieval of information.

[10] The questionnaire accompanied the home tour studies. It is not considered in the following analysis.

In general, the function of expectations is to guide effective behavior by providing a "shortcut" in mental processing (Endsley, Bolté, & Jones 2003; Roese & Sherman, 2007). Analyzing every situation anew would be very time consuming and require a lot of processing resources. That is why expectations can save processing time in many situations (Roese & Sherman, 2007); for example, people expect that the action of pushing the light switch turns on the light and do not have to think about this over and over again. However, other expectations are not quite as accurate (see, for example, Neuberg, 1996). This is not surprising when we think about the sources of expectations: stereotypes and prejudices, third-party hearsay, observations of others behaviors that have been constrained without us being aware of the constraints. Because expectations guide attention and how the person perceives information, false expectations can lead to misinterpreted data and can, therefore, negatively influence SA (Endsley, Bolté, & Jones 2003, see Section 2.1.1.2).

To conclude, in order to accelerate mental processing, expectations have to deliver information accurately and efficiently, i.e., fast and with little processing effort (Roese & Sherman, 2007). Thus, the expectations have to support the mental construction process (see Section 2.1.1.2). An expectation-driven construction process is described in Darley and Fazio (1980) with respect to a social interaction sequence between two people (see Figure 2-4).

The figure shows that the authors differentiate between perceiver and target person. However, they stress that both interlocutors can take both roles at any time and the process does not have to be symmetric. The central aspect of this sequence is the construction of the situation. In the construction process, expectations are important because they guide information gathering top-down and, thus, influence event processing. Moreover, they provide structure and meaning for the interpretation of the gained information.

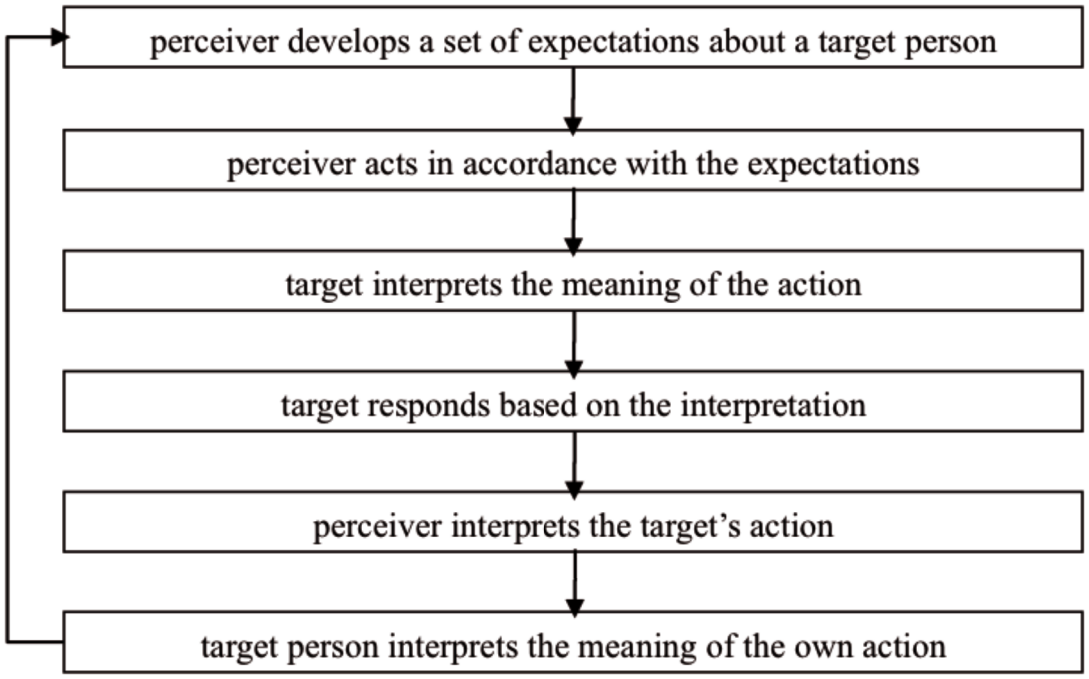


Figure 2-4. Expectation-driven construction process
(Darley & Fazio, 1980)

Based on this idea of constructivism, Ross and Conway (1986) developed a *model of personal recall*. The model emphasizes the role that response-expectancies play in the reconstruction of past events. It highlights that personal recall only depends on construction processes. However, the model ignores the fact that sometimes recall of the past is highly accurate and not always biased by inaccurate expectancies (Neuberg, 1996). To overcome this shortcoming, the *model of expectancy-guided retrieval* was developed (see Hirt, 1990; Hirt, Lynn, Payne, Krackow, & McCrea, 1999). This model describes retrieval from memory as based on several sources, i.e., information about the present, expectancy regarding stability and change between past and present, and episodic memory trace of the original information. In other words:

> *[…], the proposed retrieval process involves retrieval of the original information anchored at the outcome (the benchmark) and guided by one's expectancy." (Hirt, 1990, p.949)*

The weighting of the episodic memory trace, on the one hand, and the expectancy, on the other hand, determines the accuracy of the recall. It depends on the accessibility and the strength of the original memory trace, the motivation at retrieval, and on the assumptions on how the mind works (lay people usually assume that memories are correct and that their attitudes and beliefs do not change over time) (Hirt, 1990). If expectations and not episodic memory traces are the main source of recall, they can lead to inaccurate inferences if being incorrect, biasing information collection, or overruling consideration of information altogether (Sears, Peplau, Freedman, & Taylor, 1988).

The theory that has been presented on function and processing of expectations so far leads to the following implications for the model presented below. The function of expectations is to save time and processing resources. However, the price might be that our perceptions are not accurate, even though expectations can change and improve over time because they are part of the construction process. In other words, the users could form inaccurate expectations about the robot at first but might adapt them over time.

In fact, expectations change whenever they are retrieved from memory. They are an inherent part of our dynamic memory and strongly reflect our experiences. In other words, expectations are case based (Schank, 1999). Looking at it that way, the dynamic memory is a learning system. According to Schank (1999), learning means to alter the memory in response to experiences. Noticing and recalling mismatches to generalizations enables learning because it helps to improve outcome predictions. Having better predictions helps people to better cope with the environment. If someone encounters a deviation from an expectation he or she will most probably be reminded of episodes that are relevant for this deviation if these are stored and labeled properly. Schank (1999) calls high-level structures under which memories are stored TOPs (thematic organization packets). Within TOPs, three kinds of information are stored: expectational information, static information (knowledge of the state of the world that describes what is happening when a TOP is active), and relational information (links of TOPs to other structures in memory). Once a TOP has been selected, we begin to generate expectations. Expectations are generated by using appropriate indices to find episodes organized by that TOP.

In most cases, many memories are stored within one TOP. That is why indices are needed to find a particular memory.
More than one TOP can be active at a given time, which is reasonable because people want to be able to apply memories that are about the kind of goal they are pursuing as well as others that have nothing to do with the particular goal. Cross-contextual learning is enabled by breaking apart experiences and then decomposing them with the help of TOPs when remembering. For this process, memory structures have to be linked in the brain. According to Schank (1999), information about how memory structures are ordinarily linked in frequently occurring combinations is held in MOPs (memory organization patterns). MOPs are both memory and processing structures.

> *"The primary job of a MOP in processing new inputs is to provide relevant memory structures that will in turn provide expectations necessary to understanding what is being received. Thus MOPs are responsible for filling in implicit information about what events must have happened but were not explicitly remembered." (Schank, 1999, p.113)*

TOPs and MOPs are the structural basis of memories and information in memory (Schank, 1999). All these structures need to have the following basic content:

- a prototype
- a set of expectations organized in terms of the prototype
- a set of memories organized in terms of the previously failed expectations of the prototype
- a characteristic goal

If a situation deviates from the prototype, expectations are failed. As was mentioned above, this enables learning because with the help of TOPs it supports to improve outcome predictions. This idea is central in the model presented below because the aim of the user studies that it will be applied to is to find differences between the prototype of the designed interaction and the mental representations of the users and to identify behaviors that are connected to (dis-) confirmation of their expectations. What happens in the case of expectation disconfirmation will be discussed in the following section.

2.2.4 Violation of expectations

Most expectations are confirmed since behavior is constantly based on simple expectations as the example of pushing the light switch to turn on the light shows. The confirmed expectations are usually processed heuristically and, thus, with little effort. Then again, heuristically processed information is more likely not to lead to disconfirmation. However, under certain circumstances disconfirmation of expectations occurs. Consequences of disconfirmation of expectations are important for the following analysis because robots often involuntarily violate the users' expectations and the aim is to cope with this.

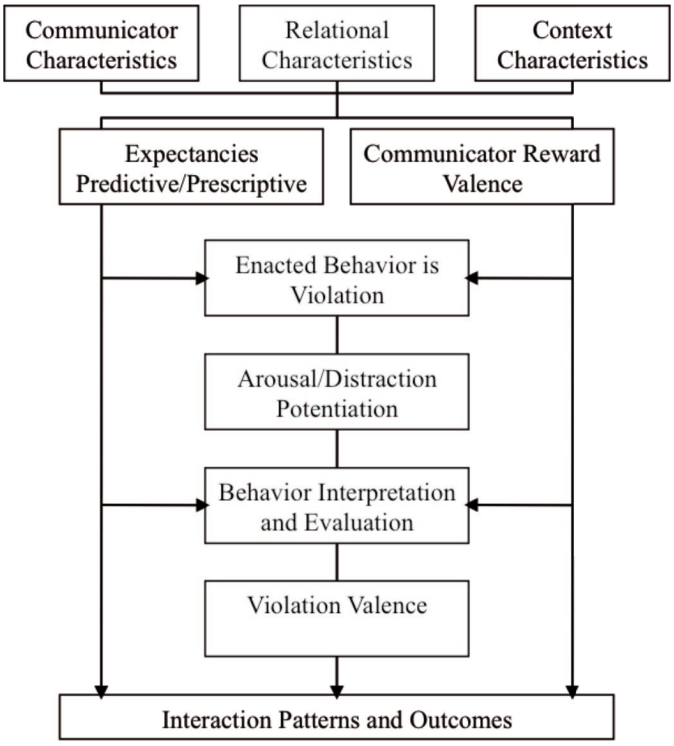


Figure 2-5. Expectancy violations theory
(Burgoon, 1993, p.34)

Here the approach presented in Burgoon (1993) seems to be helpful. As Figure 2-5 shows, Burgoon (1993) has identified three main factors influencing expectancy violation: communicator characteristics, relational characteristics, and context characteristics. The *communicator characteristics* capture salient features of the interlocutors of the interaction such as appearance, demographics, etc. The "10-arrow model for the study of interpersonal expectancy effects" (Harris & Rosenthal, 1985) is related to the communicator characteristics. It distinguishes variables that moderate or mediate expectancy effects between communicators.
Moderator variables are preexisting variables like sex, age, and personality. They influence the size of the interpersonal expectancy effects. Mediating variables describe behaviors by which expectations are communicated. Mediating variables and expectancies are modified continually in the course of the interaction. In HRI, communicator characteristics have been researched in terms of robot appearance; for example, in Lohse, Hegel, and Wrede (2008) it was found that appearance determines which tasks subjects ascribe to a robot. The influence of robot appearance is further discussed in Section 2.2.6.
The *relational characteristics* explain how the interlocutors are related to each other (degree of familiarity, liking, attraction, status, etc.). One study on relationship between human and robots has been reported by Dautenhahn et al. (2005) who raised the question "What is a robot companion – Friend, assistant or butler?". Their results indicated that the subjects preferred a household robot to be a kind of butler or assistant, but not a friend. This relation highlights that the user has a higher social status than the robot and is in control of it.
Finally, *context characteristics* include information about the situation (environmental constraints, aspects of the situation such as privacy, formality, and task orientation). The relationship between context, situation, and expectations will be further discussed in Section 2.3.

According to Burgoon (1993), these three factors determine the expectancies of the interlocutors and the *communicator reward valence*. The communicator reward valence is a measure of how rewarding it is to communicate with a certain person. It is composed of physical attractiveness, task expertise and knowledge, status, giving positive and negative feedback, and other factors. Communicator reward valence affects the interaction because it determines the valence of expectancy violations. Violations can be connected to conversational distance, touch, gaze, nonverbal involvement, and verbal utterances. They arouse and distract the perceiver and more attention is paid to the violator. Moreover, the perceiver tries to interpret the violation and evaluates it based on the interpretation, taking into account who has committed the violation. For example, the communicator reward valence is very high for a boss who is knowledgeable and gives feedback to his or her subordinates. Hence, if he or she violated an expectation, this would be interpreted differently than a violation committed by another person or by a robot with a low communicator reward valence.

Since the enacted behavior might be more positive or negative than the expected behavior, the valence of a violation can be either positive or negative. At first sight, one might suppose that the most promising strategy in interaction is to always confirm the expectations. However, some deviations might lead to better results. Two experiments by Kiesler (1973) have demonstrated that in the case of high-status individuals disconfirming may show better results and seem more attractive. On the other hand, low-status individuals are expected to act according to normative expectations and are then rated as being more attractive. As was mentioned before, Dautenhahn et al. (2005) found that the user preferred robots that had a lower status. Hence, the robots should behave in accordance with the expectations of the users and not violate them. The direction that research has taken indicates that this is what robot designers try to achieve (for example, Dautenhahn et al., 2006; Hüttenrauch, Severinson-Eklundh, Green, & Topp, 2006; Koay, Walters, & Dautenhahn, 2005; Pacchierotti, Christensen, & Jensfelt, 2005; Walters, Koay, Woods, Syrdal, & Dautenhahn, 2007). However, often the robots involuntarily violate expectations. This connects to Schank's (1999) book where he states that people can violate our expectations because they intend to or because of some error (motivational vs. error explanation). Often the reason is hard to tell. Schank (1999) has suggested reasons why people do not do what we expect:

- misperception of the situation/different perception of optimum strategy
- different goal
- lack of resources
- disbelief (not believing they should do what you thought they should do)
- lack of common sense (not knowing what to do, lack of ability)

In most situations first *error explanations* are considered. If no error can be found, the assumptions about the person's goals could be erroneous. However, determining these goals is highly speculative. If it is assumed that people knew exactly what they were doing and no error was made, a motivational explanation has to be sought. For robots, it can be assumed that the violations are mostly based on errors (unless experiments are designed to explicitly violate user

expectations). Such errors occur if the robots fail to perceive the situation correctly and lack common sense. Even worse, robots do not notice that they have violated expectations. Here the question arises of how humans recognize that this has happened.
In her model, Burgoon (1993) has not specified a mechanism that detects expectancy violations. However, Roese and Sherman (2007) have done so and they term this mechanism *regulatory feedback loops*. In the loops, the current state of a situation is compared to an ideal or expected state:

> *"Behavior control therefore requires continuous online processing in the form of continuous comparison, or pattern matching, between the current state and the expected state. Very likely in parallel to this comparative process is the conceptually similar online comparison between the current state and recent past states." (Roese & Sherman, 2007, p.93)*

The regulatory feedback loop is an online mechanism. It is triggered when people try to validate their hypothesis of the world (at least for subjective expectancies) and search for more information. Validation processes occur especially in cases of expectation disconfirmation, for example, in the case of a robot not answering when greeted.
Disconfirmation has certain consequences. If an expectation is disconfirmed, it usually becomes less certain. Moreover, disconfirmed expectations become more explicit, salient, and accessible and are, therefore, more likely to be tested in the future (see also Section 2.2.1).
In Schank's (1999) words the main question based on expectation failure that has to be asked is what changes have to be made to the memory structure responsible for this expectation? He describes the following procedure: (a) first, conditions that were present when the failure occurred have to be identified and the question has to be answered if the current context differs from situations in which the expectation was confirmed, (b) second, one has to determine whether the failure indicates a failure in the knowledge structure itself or a failure of specific expectations in that structure in order to decide what aspects of the structure should be altered. In other words, disconfirmation of expectations causes more careful processing, because humans try to make sense of the actions. Sense making includes three classes of activity: *causal attribution* to identify the cause of a particular outcome, *counterfactual thinking* to determine what would have happened under different key causal conditions, and *hindsight bias* which focuses on the question of whether the outcome was sensible and predictable (Roese & Sherman, 2007). *Attribution* is the activity that has been most deeply discussed in the literature. It is the process by which people arrive at causal explanations for events in the social world, particularly for actions they and other people perform (Sears, Peplau, Freeman, & Taylor, 1988, p.117). In this context, expectations about the future are influenced by past experiences; for example, future success is expected, if past success is attributed to ability. Thus, if the users in HRI assume that the robot has done something because they have the ability to get it to do so and the action did not occur accidentally, they probably expect that the robot would react in the same way again in a future situation. However, if the expectation is disconfirmed, the sense making activities may result in ignoring, tagging, bridging (the discrepancy is explained away

by connecting expectancy and event), and revising (in contrast to bridging involves changes at a foundational level) (Roese & Sherman, 2007). Which consequence results depends on "the magnitude of the discrepancy between expectancy and outcome and the degree of complexity or sophistication of the underlying schemata basis of the expectancy" (Roese & Sherman, 2007, p.103). Small discrepancies will more likely be ignored, moderate ones will lead to slow and steady expectancy revision, and large discrepancies lead to separate subcategories. Especially if similar deviations are encountered repeatedly, at some point these form a new structure. When a new structure is used repeatedly, the links to the episodes it is based on become harder to find (Schank, 1999).

Whatever change is made to memory structures, only the explanation of an expectation failure protects us from encountering the same failure again (Schank, 1999). Explanations are the essence of real thinking. However, people are usually content with script-based explanations if these are available (for more information about scripts see Section 2.2.5.3).

While discrepancies may lead to changes in memory structure, perceivers may also confirm their inaccurate expectancies. According to Neuberg (1996), perceivers may confirm their inaccurate expectancies in two primary ways: by creating self-fulfilling prophecies (target behavior becomes consistent with the expectancies) and by exhibiting a cognitive bias (target behavior is inappropriately viewed as being expectancy-consistent). Thus, the perceivers impose their definition of the situation on the target, or even affect their behavior, because they want to confirm their own expectations.

Olson, Roese, and Zanna (1996) have also investigated affective and physiological consequences of expectancies (for example, placebo effect) and expectation disconfirmation. Affect usually is negative when expectancies are disconfirmed, an exception being positive disconfirmations like a surprise party. In general, an expected negative outcome is dissatisfying, an unexpected negative outcome even more so. The other way around a positive outcome is even more satisfying when being unexpected.

To conclude, the goal of HRI should be that the robot does not disconfirm the users' expectations, especially not in ways leading to negative outcomes. Unfortunately, it is difficult to achieve this goal. Therefore, it is important to take into account the consequences of disconfirmation which depend on the characteristics of the human and the robot, the relationship between the interactants, and the context. One part of the model presented below is to determine how the users make sense of the disconfirmation of their expectations by the robot and what are the results of the sense-making process. Knowledge about this will help to design the robot in a way to better contribute to the correct explanation of why the disconfirmation occurred or to avoid the disconfirmation in the first place.

2.2.5 Expectation-related concepts

The notion of expectation is associated with many other terms. In this section, expectations are contrasted with belief, schema, script, and attitude, in order to further specify the expectation concept.

2.2.5.1 Beliefs

LePan (1989) has distinguished expectations from other future-oriented feelings such as believing:

> *"A belief, to begin with, may fly in the face of all logic, of all rational analysis of the probability or improbability of a given occurrence taking place. Expectation, on the other hand, is firmly grounded in a rational assessment of probabilities: our reason may be, and often is faulty, but it is nevertheless reason rather than faith on which expectations are grounded." (LePan, 1989, p.73)*

Accordingly, the terms expectation and belief are distinguished by their certainty and by the fact that expectation in contrast to belief is based on reasoning. However, Roese and Sherman (2007) in their definition of expectancies state: "Expectancies are *beliefs* about a future state of affairs". So seemingly both positions disagree. Olsen, Roese, and Zanna (1996) clarify the discrepancy as follows:

> *"All beliefs imply expectancies; that is, it is possible to derive expectancies from any belief" (p.212)*

According to this opinion, beliefs are the basis for expectancies. Olson, Roese, and Zanna (1996) give one example to demonstrate this relationship – a belief would be that fire is hot; the expectancy which can be derived from the belief is that people burn themselves if they hold their hand in a flame. According to the authors, not all beliefs are expectancies, but expectancies are a type of belief about the future. The type of interest for the expectancy concept is implications of beliefs that are not yet verified.

2.2.5.2 Schemas

Schemas are often mentioned in connection to expectancies. Aronson, Wilson, and Akert (2003) define schemas as:

> *"mental structures people use to organize their knowledge about the social world around themes or subjects and that influence the information people notice, think about, and remember" (p.59).*

This definition shows that expectancies and schemas are similar in that both influence information processing and memory, a view that has also been supported by Hirt (1999). Sears, Peplau, Freedman, and Taylor (1988) describe the content of schemas:

> *"a schema is an organized, structured set of cognitions about some concept or stimulus which includes knowledge about the concept or stimulus, some*

> *relations among the various cognitions about it, and some specific examples." (p.101)*

According to the definition, schemas include knowledge, relations, and examples of a certain concept. The accessibility of schemas increases if they are primed. Priming is "the process by which recent experiences increase the accessibility of a schema, trait, or concept" (Aronson, Wilson, & Akert, 2003, p.64). This is also true for expectations.
Some authors have explicitly described the relation between schemas and expectations. As Sears and colleagues (1988) write, schemas contain expectations for what should happen. Olson, Roese, and Zanna (1996) describe the relationship between schemas and expectancies as follows:

> *"Schemas are mental structures that include beliefs and information about a concept or class of objects […]. One of the primary functions of schemas is to allow the generation of expectancies in situations where full information is not available." (p.212)*

From this point of view, like beliefs, schemas are a source of expectancies, but not expectancies themselves. Hirt (1999) supports this argument by stating that people's expectations are based on their schemas. Feather (1982b) has differentiated schemas and expectations by the degree of specificity. According to the author, expectations are much more specific than schemas, because they relate to particular situations and responses. Therefore, they are more closely related to scripts than to schemas,

> *"because the specific content of both expectations and scripts is a fundamental aspect of their definition and both concepts are concerned with the structure of events within defined situations and the implications of particular actions within those situations." (Feather, 1982b, p.407)*

2.2.5.3 Scripts, scenes, and scriptlets

While Feather (1982b) admits that the content of both scripts and expectations is similar, he also argues that scripts are more extensive concepts that include whole sequences of events whereas expectations can be based on single events. A set of specific expectations might result in a script.
Schank and Abelson (1975) defined the term script as "a structure that describes an appropriate sequence of events in a particular context" (p.151). A script consists of slots and the content of one slot affects the content of other slots, because all slots are interconnected in causal chains. Actions that do not match the causal chain are deviations from the script. Deviations can be *distractions*, i.e., interruptions by another script; *obstacles*, i.e., a normal action is prevented from occurring; and *errors*, i.e., the action is completed in an inappropriate manner. Each script includes different actors. Each actor has an own point of view on the script, for example, a

script of an HRI situation of the person interacting with a robot differs from the script of a bystander. This view is supported by Anderson's (1983) definition:

> *"[…], we frequently think about our own actual or potential behaviors; that is, we create behavioral scenarios (**scripts**) in which we are the main character." (Anderson, 1983, p.293)*

In this definition, Anderson argues that scripts always focus on our own behavior. He furthermore writes that scripts influence expectations and intentions, interpretations of situations and behavior. Thus, similar to schemas, scripts trigger expectations.
Later, Schank (1999) has revised his original notion of scripts which has been cited above (Schank & Abelson, 1975). As he admits, the original approach of him and his colleagues was highly modular, breaking the task of understanding into discrete and serially executed components.

> *"In those days, scripts were intended to account for our ability to understand more than was being referred to explicitly in a sentence by explaining the organization of implicit knowledge of the world we inhabit. […] In general, we ignored the problems of the development of such structures, the ability of the proposed structures to change themselves, and the problems of retrieval and storage of information posed by those structures. We concentrated instead on issues of processing." (Schank, 2005, p.6f.)*

Thus, in their first definition, scripts were referred to as a data structure that was a useful source of predictions, based on repeated experience. However, this use of the term was not in agreement with their theory. Schank (1999) underlines this with the following example: few people have been in an earthquake; however, everybody has knowledge about such events and understands stories about them. Is this knowledge then represented in scripts? Based on this example, Schank (1999) distinguishes scripts and other high-level knowledge structures with regard to abstraction or generalization. Only more general information can be used in a wider context than the one it was acquired in. In the original notion of the concept, too much information was stored specifically in a script. In contrast, it should be stored generally to allow for generalizations.
Moreover, the original definition viewed scripts as a passive structure that cannot be changed easily. In contrast, in the revised definition Schank stresses that scripts are an active memory structure that changes in response to new input. From this new point of view, no complex structures of complex events do exist in memory, but the scripts and the expectations are distributed in smaller units that have to be reconstructed.
Schank (1999) introduces two terms that are needed to fill out the new definition, i.e., scenes and scriptlets. *Scenes* are one kind of memory structures that provide a physical setting serving as the basis for reconstruction of an event. The most important scene has to be found for interpretation. A scene is likely to be ordered by using different contexts.

> *"Scenes […] can point to different scripts that embody specific aspects of those scenes. Scripts can also hold memories that are organized around expectation failures within that script. According to this theoretical view, a script is bounded by the scene that contains it. Thus, scripts do not cross scene boundaries." (Schank, 1999, p.113);*

Scriptlets are copies of a scene with particulars filled in. In case of expectation failure a new scriptlet is copied which includes the necessary changes to the scene. This allows for knowledge transfer across contexts. Scriptlets are acquired in learning by doing. They tend to decay in memory if they are not used. Schank (1999) differentiates three groups of scriptlets: cognitive, physical, and perceptual scriptlets. All scriptlets are static; for example, if a person acted based on a scriptlet only, she would have to order the same thing every time when eating out at a specific restaurant. Thus, some scriptlets should become scenes over time. Only if the scriptlet is transformed into a scene, the person is able to order whatever is on the menu.

2.2.5.4 Attitudes

Another concept related to expectations is *attitudes*. Attitudes are enduring evaluations of people, objects, or ideas (Aronson, Wilson, & Akert, 2003). A main aspect in the traditional view of attitudes is that they are enduring. However, this assumption has been challenged based on the constructivist idea that people construct attitudes from the most accessible and plausible information like their own behavior, moods, and salient beliefs about the attitude object. The construction process is influenced by the social context, meaning that they depend on what information comes to mind, how it is evaluated and used for judgment (Bohner & Wähnke, 2002).

On the other hand, Wilson and Hodges (1992) have found that many attitudes are stable and changes only occur if they are less stable, if people are less involved, or if they have an inconsistent structure. However, all attitudes are influenced by goals, moods, bodily states, standards, and appropriateness (Bohner & Wähnke, 2002). They can have cognitively, affectively, and behaviorally based components (Aronson, Wilson, & Akert, 2003) and they affect thinking on all levels: (a) attention, encoding, exposure; (b) judgment and elaboration; and (c) memory (Bohner & Wähnke, 2002).

Concerning their operationalisation, attitudes have been calculated as the sum of expectancy x value products (the expectancy of the behavior having certain consequences multiplied with the subjective value of these). Both, expectancy and value are basic variables determining motivation tendencies that build the basis of what we finally do (Heckhausen & Heckhausen, 2006). However, many more factors influence the interaction. For example, studies have provided a lot of evidence that performance at a task is affected by initial performance (Feather, 1982a). That is why after a failure the expectancy is low and after a success it is high. This is only one factor influencing behavior. There are many more variables, for example, changing expectations during the task performance, individual differences, and characteristics of the task. Therefore, attitudes often do not predict behavior accurately but situational variables have a greater influence than personal variables (Feather, 1982a). That is why for the following analysis a

situation- and expectation based model is developed. However, expectancy-value theories are an important part of motivation research. Examples of expectancy-value models, for example, Atkinson's Theory of Achievement Motivation, are discussed in Feather (1982b).

To conclude, expectations are generated based on schemas and scripts. While schemas are more general, scripts are as specific as expectations. However, both schemas and scripts of actions are *descriptions* of components of these actions, whereas expectations are *implications* and attitudes are *evaluations* of these descriptions.

2.2.6 Empirical studies concerning expectations in HRI

This chapter has already drawn some connections between expectations and HRI. This section presents the sparse research that actually exists on expectations in HRI. Some studies have been conducted with respect to normative expectations. Khan (1998) reported a survey with 134 Swedish subjects which suggested that certain tasks were preferred for intelligent service robots: robots should help with chores like polishing windows, cleaning ceilings and walls, cleaning, laundry, and moving heavy things. However, they were not wanted for tasks like babysitting, watching a pet or reading aloud. Altogether, 69% of subjects found the idea of a robot as assistant positive, 44% found it probable, 23% frightening, and 76% thought that the concept was useful (Khan, 1998).

Arras and Cerqui (2005) have supported the idea that the attitude towards robots depends on their tasks. In a survey with more than 2000 participants at the Swiss expo, 71% exhibited a neutral attitude towards robots, meaning that they neither rejected them nor were completely positive. This is due to the fact that the result is based on the mean of attitudes toward robots used for different tasks. As mentioned above, some tasks were welcome while others were not.

These findings have recently been supported by Ray, Mondada, and Siegwart (2008) in their paper titled "What do people expect from robots?". In a questionnaire study on people's perception of robots with 240 participants, the authors found a positive attitude towards robots in general. But they also reported that people preferred robots mainly as helpers in daily life and not as caregivers in whatever context. They noted that people expressed their fear about robots replacing humans and about robots becoming too autonomous. Ray, Mondada, and Siegwart's (2008) findings revealed that robots should not be humanoids or look like living beings. Nevertheless, people prefer natural channels for communication with robots, for example, speech.

Although these studies ask for users' expectations, they do not focus on expectations in interaction but rather on evaluation of general attitudes towards robots. These have also been in the focus of research with the NARS (negative attitudes towards robots) scale, which is based on the concept of computer anxiety (for example, Bartneck, Nomura, Kanda, Suzuki, & Kato, 2005; Bartneck, Suzuki, Kanda, & Nomura, 2006; Nomura, Kanda, & Suzuki, 2004; Nomura, Kanda, Suzuki, & Kato, 2004). However, researchers using this scale barely identified a link between attitudes and behavior. This might in part be due to the fact that there is no general attitude towards all robots but the attitudes strongly depend on the tasks of the robots, as has been mentioned before.

First inferences about the tasks of a robot are made based on its appearance. Robot appearance has a major influence on the assumptions people form about applications and functionalities, i.e., behaviors of robots (Goetz, Kiesler, & Powers, 2003). Appearance influences the first impression and expectations we have of somebody or something. “With robots first impressions count. In a few minutes, any user will have made his or her first opinion about the object” (Kaplan, 2005, p.4). This usually already happens before the interaction starts and influences how it will proceed. Because of the appearance of a robot, users generate expectations about its abilities; for example, if a robot has a camera that reminds them of eyes, users expect that the robot can see. The interaction will only be enjoyable if the actual functionality matches or exceeds expected functionality (Kaplan, 2005). That is why “the design should convey clear message about the type and context of usage of the robot. More importantly it should trigger the right kind of expectancies” (Kaplan, 2005, p.4).

The expectancies are influenced by previous interactions with humans, animals, and machines. Because of these experiences, particular perceptual features of the robot trigger schemas of interaction (see Section 2.2.5.2). Then again, expectations are triggered by the active schemas. That is why current research states that the appearance has to support the correct estimation of the robot’s real competencies by the users. The better the users’ estimation, the less will they be disappointed during the interaction with the robot (Kaplan, 2005).

Fong, Nourbakhsh, and Dautenhahn (2002) have defined four broad categories of social robots with respect to their appearance: anthropomorphic, zoomorphic, caricatured, and functionally designed robots. It has been shown that the more human-like the appearance of a robot is, the more people attribute intentions to it within a Prison Dilemma Game task (Hegel, Krach, Kircher, Wrede, & Sagerer, 2008). On this background, two studies researched which tasks people attribute to robots based on their appearance (Hegel et al. 2007; Hegel, Lohse, & Wrede, 2009; Lohse et al., 2007; Lohse, Hegel, & Wrede, 2008).

In the first online survey, the 127 participants were presented with four robots in random order: BARTHOC, iCat, AIBO, and BIRON (see Figure 2-6). They received very basic information about the functionalities of each robot. Next to the descriptions, videos were displayed on the screen. In about three seconds they showed few movements of each system to give an impression of the robots’ appearance. Based on this information, the participants’ task was to propose applications for the robots. 570 applications were named and then categorized into the following 13 groups (for details see Lohse et al., 2007). *Healthcare* refers to robots used for therapy (for example, autism therapy) and as support for sick or old people. This category also includes *Caregiver* robots that are used to watch old or sick people when nobody else is around. *Companionship* consists of all robots that keep company. The purpose of *Entertainment* robots is to entertain their users and to be a pastime. They are not built to have a psychological impact. The same is true for *Toy* robots that are mainly used for playing. Most robots currently being sold for domestic usage belong to this category. Another application is *Pets*. It implies that the user shows responsibility for the robot. Pet robots are animal-like in appearance and functionalities and might take the place of a real pet. *Personal assistant* or *Interface* describes robots used as butlers, organizers, or interfaces. The category includes robots for cleaning and other household chores. *Security* applications concerns robots used for surveillance, military tasks,

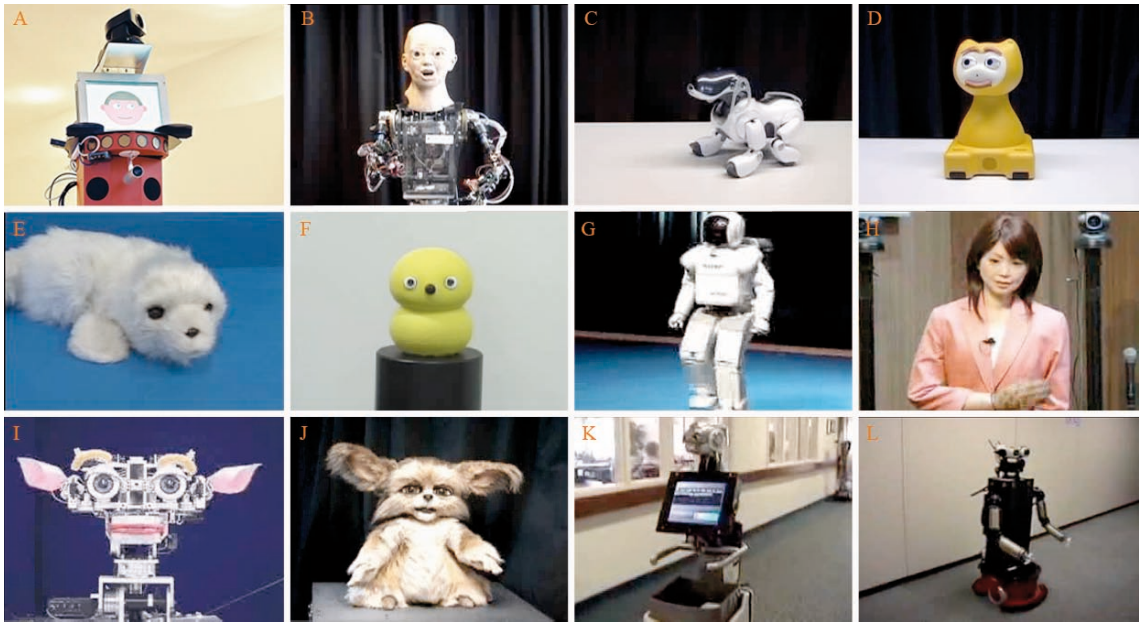


Figure 2-6. Robots of the appearance study
A) BIRON, B) BARTHOC, C) AIBO, D) iCat, E) Paro, F) KeepOn, G) Asimo, H) Repliee Q2, I) Kismet, J) Leonardo, K) Pearl, L) Robovie

exploration, tasks that are dangerous for humans (for example, minesweeping), and for protection. Another category includes robots that *teach* certain tasks or abilities. The robots in this case are supplements to real teachers especially when learning languages. *Transport* robots are useful for all kinds of fetch and carry tasks. *Business* robots are receptionists, sales robots, or robots used for representation. Finally, *Public assistants* are guides (for example, in museums), information terminals or translators. Different users employ these robots usually only for a short time at once. The applications named most often were Toy (105), Public Assistant (90), and Security (77).

A second survey (Hegel, Lohse, & Wrede, 2009) was conducted with 183 subjects to verify the categories obtained in the first study with an international sample and a higher number of robots. This study included eight more robots: KeepOn, Kismet, Leonardo, Robovie, Repliee Q2, Asimo, Paro, and Pearl (see Figure 2-6).

The participants were provided with the 13 application categories that were determined in the first survey and their task was to rate how suitable each robot was for the application. In general, the participants expected that the robots were most appropriate for the applications entertainment (3.4) (all means on a scale of 1 [not at all] to 5 [very much]), toy (3.3) and research (3.1). Toy was also the most popular category in the first study.

Altogether, the two studies have shown that the participants readily attribute tasks to robots based only on a first impression of their appearance. Judgments of appropriateness for different tasks are made within few seconds. But the studies do not only provide information on which robots seem suitable for certain tasks. They also allow to infer whether all robots that appear suitable for a certain application have something in common, or in other words, if there are basic expectations the robots have to fulfill on the first sight to seem suitable for a specific application. One such basic requirement is the agreement of certain tasks with certain basic types of appearance; for example, a high degree of human-likeness for tasks that are high on

social interaction (for example, care giving, teaching). In contrast, companion, pet, toy, and entertainment applications do not imply a necessity for human-like appearance, but rather animal-like appearances are preferred. This expectation might have been influenced by the fact that commercial entertainment and toy robots nowadays often have an animal-like appearance.
Appearance may also communicate specific functions and abilities of the robot such as mobility and ability to carry objects. This functional aspect of appearance was found to be very important with respect to the participants' expectations. For example, only robots that appeared mobile were suggested for transport tasks. But not all robots were perceived the same by all participants, as has been shown for Aibo that was rated as being highly functional *or* highly animal-like. Thus, expectations do not arise from the physical appearance alone, but from the perception of the appearance by the participants.
As the other approaches presented in this section, the two studies served to determine the expectations the people start the interaction with. But these cannot be expected to reliably explain the users' behavior during the interaction, for example, one could assume that if the robot is designed to serve as a pet and also looks like one, the user will produce behaviors that pet owners use for communication with their pet. However, it must be taken into account that the robot does not act like a real pet and does not have the same abilities. Once the users notice these short-comings, they will probably change their behavior and it cannot be predicted any-more. Even if the users stuck to pet owner behavior, one could argue that a reliable prediction is not possible because it depends too much on the situation and the individual expectations of a person. That is why a model of HRI that takes these factors into account is needed.

2.3 The notions of situation and expectation in HRI

In the following, assumptions that could be retrieved from the theory are summarized. On this basis, a model of the influence of situation and expectation on HRI is proposed. The main assumptions are:

1. the HRI situation exists as a physical social situation that can be described on different levels of specificity
2. the physical situation restricts the behavior repertoires of the interactors
3. the users perceive the situation based on their knowledge and the information that they take into account (i.e., the context that they consider)
4. the users choose their behaviors based on their perception of the situation
5. the users' behavior allows for implications of their perception of the situation
6. the behavior repertoires of the users and the robot differ and the interaction is asymmetric
7. the perception of the situation changes throughout the interaction
8. the expectations change throughout the interaction (in a first-contact situation the expectations are first based on HHI and become more target-based, i.e., specific expectations about the robot are developed)
9. if expectations are supported, their likelihood increases; if expectations are violated, their likelihood decreases
10. the expectations allow the users to improve their outcome predictions, i.e., to learn

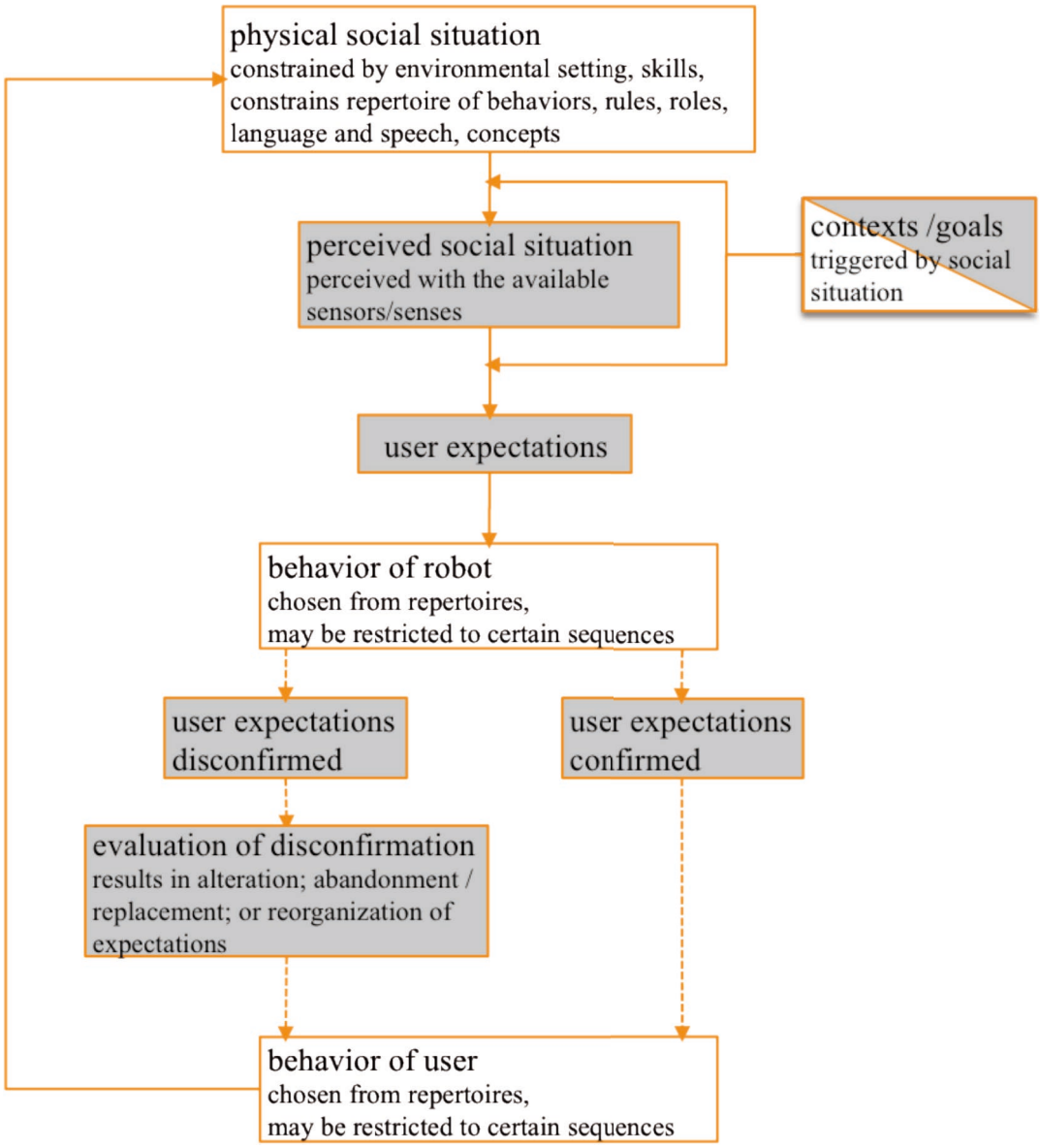


Figure 2-7. Model of situation and expectations in HRI

These assumptions are the basis for a model to describe HRI in a framework of social situations and expectations (see Figure 2-7). The model analyzes the physical situation and the behaviors (white boxes because these can be observed) in order to infer how the user perceived the situation, what expectations the user had, and whether these were confirmed or disconfirmed (grey boxes because these cannot be observed but need to be inferred). The model takes into account that each interactor has an individual perception of the situation. It focuses on the interaction between one user and one robot.

The model is based on the first assumption that a *physical social situation* exists (assumption 1) (Goffman, 1961; Magnusson, 1981a; Rohlfing, Rehm, & Goecke, 2003; Wagner & Arkin, 2008). This situation is social because the user and the robot are present and both monitor and influence each other (Goffman, 1961). The physical social situation describes what really exists, such as agents, objects, time, and space (Magnusson, 1981a; Rohlfing, Rehm, & Goecke, 2003). It is constrained by the environmental setting and the skills of the interactors (Argyle, Furnham, & Graham, 1981; Smith, 2005). Moreover, the physical situation constrains the repertoire of

behaviors of both interactors, the rules that apply within the situation, the roles of the interactors (which can change as the situation progresses), the language and speech that is appropriate for each interactor, and the concepts that help the interactors to understand the behavior of the counterpart and to plan their own behavior (assumption 2) (Argyle, Furnham, & Graham, 1981). These factors are here regarded as part of the physical situation because they are to a high extent based on norms that exist within society and are, thus, also accessible for observers. The behavior repertoire that the user chooses from is usually much bigger than the robot's behavior repertoire because the user has more knowledge than the robot (assumption 6).

The physical situation is perceived by the user and is then called *perceived social situation* (assumption 3) (Craik, 1981; Magnusson, 1981b). The perception process is strongly influenced by the *contexts* that the user considers when perceiving the situation (Schank, 1999). Here, the term context refers to dynamic repositories of knowledge gained in previous situations (see Section 2.1.2). The model assumes that part of the contexts is known to the observer because they are actively evoked, for example, by the setting (for example, the robot apartment) that reminds the user of similar settings, by the social interaction with natural modalities that aims to remind the user of social HHI, and by the tasks (here teaching) that remind the user of similar tasks and imply certain goals. However, the user might also take other contexts into account that have not been consciously evoked by the observers. Part of these contexts can be inferred through cues provided by language, perception, and attention of the user (assumption 5) (for example, Fischer, 2000), because the user chooses the behavior based on his or her perception of the situation (assumption 4) (see for example, Magnusson, 1981b; Ross & Nisbett, 1991).

The user's perception of the situation changes throughout the interaction (assumption 7) (Bierbrauer, 2005; Rohlfing, Rehm, & Goecke, 2003; Rotter, 1981, Smith, 2005). Based on the dynamic perceived situation and contextual knowledge, the user develops *expectations* towards the robot. These expectations are here seen as dynamic because they evolve with the situation, i.e., they change with the perception of the situation and the knowledge that is gained during the interaction (assumption 8) (Heckhausen, 1977; Hirt, 1990; Schank, 1999). Once the *robot* performs a *behavior* from its *repertoire*, the expectations are confirmed or disconfirmed. If they are confirmed, their likelihood increases and if they are disconfirmed, it decreases (assumption 9) (Roese & Sherman, 2007). Behaviors of the robot here include verbal behaviors, movement of the camera, and spacing behaviors. If the users' expectations are disconfirmed, they usually *evaluate the disconfirmation* trying to find an explanation for it. Evaluation of disconfirmation can result in the alteration of a specific expectation, abandonment or replacement of the expectation, or reorganization of the expectation (Schank, 1999). This process can be termed *learning* (assumption 10) (Schank, 1999). Based on his or her current expectations, the *user performs a behavior* from the individual repertoire which might follow a certain sequence. *Behaviors* of the user include verbal behaviors, gestures, facial expressions, gaze, and spatial behaviors. They are constrained by the context, in space and time, physically and socially (for example, Argyle, Furnham, & Graham, 1981). It is important to keep in mind that behaviors need to be differentiated from inferences; for example, the user clicking a button is a behavior, whereas the user being confused is an inference (Hackos & Redish, 1998).

As can be seen in Figure 2-7, the model is cyclic and the process starts over. The physical social situation now changes because the behavior of the user becomes a part of it. In the next turn, the user again perceives the situation and can, thus, evaluate the effects of the own behavior and adapt his or her expectations.

The model introduced here fulfills the requirement of incorporating the human and the robot, the influence of the situation and of the expectation of the user. Moreover, it takes into account that the situation is dynamic which shall make it suitable to better predict user behavior. Whether this can actually be achieved with the model will be shown in the following chapters that focus on the analyses of empirical user studies.

3 Methods and novel developments of HRI data analysis

The previous chapter has shown that the physical social situation, part of the contexts/goals, the behavior of the robot and the user can be observed and with their help the users' expectations and their perception of the situation can be inferred. In the following, methods are introduced that were developed and combined to research the observable factors of the interaction and to infer the non-observable factors.

The measurement of expectations is difficult because expectations are reactive to it (Olson, Roese, & Zanna, 1996). This means that the measurement process itself may induce expectations that would not have been generated spontaneously. That is why self-report measures alone are fallible (Feather, 1982c). To approach this problem, in the following, various sources of inference of user expectations in HRI are combined. The methods include user behavior, the physical social situation as depicted in the model, and self-report data.

In order to infer expectations from user behavior, it needs to be closely analyzed. To enable this analysis, various modalities need to be taken into account because they combine to form a coherent behavior that is full of information. The multimodal data has to be coded in a way that allows for statistical analysis. Therefore, coding schemes are developed for the different modalities (see Section 3.1). Each code describes a distinct behavior. To analyze the behaviors, SALEM (Statistical AnaLysis of Elan files in Matlab) is introduced as an efficient quantitative approach. SALEM is used in the following for the analysis of all modalities in the object-teaching study as well as in the home tour, i.e., for verbal actions, gestures, spacing behavior, and eye gaze. How the modalities are analyzed and how they contribute to the inference of expectations is described in Section 3.2. Also, possible sequences and interrelations between the modalities will be discussed.

The relation between HRI and expectations can only be researched based on situated interaction data because (1) real users have different expectations than developers and (2), as has been discussed before, the situation determines the expectations. The situation needs to be closely analyzed in order to infer expectations that are connected to it. While this analysis can be straightforward for restricted situations such as the object-teaching study where the task, the utterances of the robot, and the space were clearly defined, it becomes very complex in situations like the home tour where the behaviors of the autonomous robot cannot always be foreseen and the interactors move freely in space to complete various tasks. Therefore, a method is needed to describe these more complex situations and the problems that arise within them on the task level. In the following, the Systemic Interaction Analysis (SInA) approach is used to conduct this description of the situation (see Section 3.2.6). It takes into account the interaction level (interplay between user and robot) and the system level (robot components and their interplay) in order to determine how the robot behaviors influence the users' expectations, which user behaviors result from the expectations, and how the user behaviors in turn influence the robot behavior.

Sometimes the relation between the user behaviors cannot be determined without taking a closer look at the sequences and the timing. In these cases, the data were visualized which allowed for

a direct, qualitative comparison (see Section 3.4). Even though the users performed the same behaviors, their timing pointed to different expectations.

Although observing the interaction itself contains much information about expectations, also what happens next to the interaction can be of importance. That is why the methods include off-talk analysis (talk between the user and the experimenter during the interaction), questionnaires, and interviews (see Sections 3.5 and 3.6). The self-report data will be taken into account to further validate the inferences that are made based on the users' behavior. Section 3.7 gives an overview of all methods and summarizes their goals.

3.1 Data-driven development of coding schemes

Modalities are used differently in varying interaction situations, which is why no universal coding schemes exist in HHI that could be transferred to HRI. The main problem with universal coding schemes is that the behaviors are determined by the situations. Thus, universal coding schemes would have to be very general and abstract to be applicable for many situations and much information would be lost in the analysis process. Therefore, coding schemes need to be data-driven to actually include the behaviors that occur in a certain situation. Moreover, the coding schemes depend on the research goals that strongly influence their content and the granularity with which behaviors are coded. For these reasons, coding schemes were here developed for speech, gesture, and body orientation (see Table 3-1). These were based on the video recordings of the interaction which were coded manually. In the object-teaching task speech and gesture were of interest because these were the main modalities that the users communicated with. In the home tour speech was not investigated because the users were trained how to speak to the robot. But body orientation was taken into account as an additional modality because the robot and the user were free to move in the apartment and the orientation between them changed communicating information about the interaction situation.

Table 3-1. Overview of coding schemes

Object-teaching 1	coding schemes for speech and gesture in the object-teaching task (Section 4.1)
Object-teaching 2	refined coding schemes for speech and gesture in the object-teaching task (Section 4.2)
Home tour 1	no coding schemes
Home tour 2	coding schemes for gesture and human body orientation with respect to the robot and objects in the home tour (Sections 5.1.2, 5.1.3)

3.2 Quantitative analysis of the users' behavior repertoires

The quantitative analysis of behaviors focuses on the questions of what the users do, why they do it, and what the behavior tells us about their expectations. It is claimed here that knowing the users' expectations in certain situations, helps to predict their behavior in similar situations and to adapt the robot system to it. Since the situation strongly influences the interaction, this generalization cannot be made across situations. The following analysis will show how far, however, it can be made across situations that share the same communicational function, for example, social interaction situations and situations that evolve around the same task.

The behavior of the users is analyzed with respect to their repertoires, i.e., all behaviors that occurred with respect to the modalities were identified. These make up the users' behavior

repertoires. The users chose different elements from their repertoires based on the interaction situation, which determines what role the users are in and what rules apply. Moreover, the situational goal determines what elements are chosen. The users' behavior repertoires are further restricted by the language that the robot understands and by the setting (see Section 2.1.1.3). How exactly the setting influences the repertoire is one important question for the following analysis. If it can be found that the elements have certain functions and certain elements are used in certain situations, this knowledge could help to enable the robot in a multimodal way to identify what the user is trying to do and to better adapt to the situation. Also changes between the elements might be useful hints. If these changes occur in a systematic way, they will be referred to as sequences.

The behaviors that are analyzed in the following can be divided into verbal and nonverbal behaviors. Verbal behavior here means speech, i.e. what is said. It does not take into account how it is said. Nonverbal behavior refers to the way in which people communicate, intentionally or unintentionally, without words (Argyle, 1975). Nonverbal cues include facial expressions, tone of voice, gestures, body position and movement (spatial behavior), posture, touch (bodily contact), and gaze. Argyle (1988) also includes clothes and other aspects of appearance, smell, and non-verbal vocalizations. Each of these contains a number of variables (for example, gaze includes mutual gaze, length of glances, and amount of eye-opening). They serve to ease verbal communication, to express emotions, communicate attitudes and one's own personality traits. The meaning of non-verbal signals depends on the social setting, the position in time and its relation to other signals (Argyle, 1988).

In the following, speech and three nonverbal behaviors are analyzed: gestures, spatial relations, and eye gaze. Other modalities are not taken into account since they are not present in the type of HRI researched here, i.e., touch is not of importance. It does not have any communicative function for BIRON, because the robot does not have touch sensors and the users do not touch it in the given scenario. The same is true for smell and heat. The robot was not able to detect facial expressions and tone of voice either. Moreover, it expressed neither of these modalities deliberately. Therefore, it is assumed that the robot's behavior on other channels does not have a systematic effect on the participants' facial expressions and tone of voice and, hence, researching these modalities would not be very rewarding in the sense that the results would increase the robot's ability to better understand the situation.

For the modalities that are analyzed, the upcoming sections will target the following questions:

- What does the modality describe?
- How is the modality represented in the robot?
- Why was the modality chosen?
- Is there relevant related work in HRI concerning the modality?
- How was the modality annotated and analyzed?

3.2.1 Analysis of speech

Speech can be considered to be the most important modality in the home tour because both the human and the robot speak to establish a social relationship, to signal attention, and most importantly to complete the tasks. Natural interaction in this scenario really means that the user and the robot talk to each other like humans talk to other humans. According to Mills and Healey (2006), the function of speech in HHI can be categorized into four levels:

Level 1 - Securing Attention
Level 2 - Utterance recognition
Level 3 - Meaning recognition
Level 4 - Action recognition

If one of these levels is not secured, miscommunication can occur. The most basic level is to attain the attention of the interaction partner. Voice can be used to direct attention (Clark, 2003). Speakers use it to indicate themselves as speakers, their location as here, or the time of utterance as now. Level 2 requires that utterances as such are recognized, i.e., the listener notices the speaker's utterance. In level 3 the listener also understands the meaning of what was said. The fourth level underlines that the verbal behavior is important to recognize what people are trying to do. This is also true for the users in HRI. The content of the users' utterances indicates what task they are trying to complete and what they intend to call forth in the hearer (here the robot). What the users intend the robot to do is encoded in the *evocative* function of the utterance (Allwood, 2001). Also, the other levels have an evocative function. According to Allwood (2001), default evocative functions are continue, perceive (Level 2), understand (Level 3), and react according to the main evocative function (Level 4), where the main evocative function is the action that the user tries to achieve (for example, the robot follows when asked to do so). There are many ways to categorize utterances, for example, according to the purpose of speech (egocentric utterances, questions, influencing the behavior of others, conveying information, sustaining relationships, etc.), with respect to rewards and punishments, and with regard to the topic of conversation (for example, show solidarity, disagree, ask for suggestions) (Argyle, 1969). All the categorizations presume the interpretation of what was said.
Utterances also have an *expressive* function to express beliefs and other cognitive attitudes and emotions (Allwood, 2001). An expressive function in HRI could be that the user, by asking the robot to do something, expresses the belief that the robot is actually able to do so. The user can also express positive emotions when the interaction is going well and frustration when some problem occurs.
How speech is annotated is influenced by the question of which aspects of the expressive or evocative functions play the main role in research. For research on expressive functions the coding has to take the particularities of the speech and its intonation into account. In contrast, with respect to the evocative function it is more important to identify what was said. In the following analysis, the evocative function of speech plays the main role. What is said is more important than how it is said, because speech is mainly analyzed to determine what the users try to achieve and what tasks are in the focus of their attention. Moreover, the robot was not able to

change its intonation. All speech followed the same intonational patterns. The author is not aware of research that investigates how this influences the way the human speaks. One could hypothesize that the users adapt to the robot and intonate less. However, this question opens the door to a whole new research field which is not in focus here.

Speech in HRI

As has been mentioned before, task completion is one main aspect of the interaction in the home tour scenario. Accordingly, the annotations of the users' speech shall be used to recognize the actions that they aim to evoke. Only a small amount of data exists on how people would try to evoke actions in embodied communicational agents or robots. Kopp (2006) has presented results from how people communicated with the agent Max in the Heinz Nixdorf Museum (Paderborn, Germany). He and his colleagues analyzed what people said to the agent, i.e., which utterances they typed in using a keyboard in this public setting where no experimenter was watching and where they did not have a concrete task. The utterances were clustered for analysis according to their content. Different actions were found: greeting and farewell, insults, random keystrokes and senseless utterances, questions (concerning system, museum, checking comprehension, etc.), answers (for example, to questions for age, name), requests for attention (talk to me), and requests like shut up (which could most probably also be clustered as insults). Lee and Makatchev (2009) found quite similar results for the interaction with the roboceptionist robot where people also had to type in the utterances. The main dialog topics were seeking information (as one would do with a real receptionist), chatting, saying hello, nonsense/insult. In general, the authors found two groups of people: the ones using full sentences and the others who used keyword commands (but displayed more polite behaviors altogether). These findings were certainly influenced by the fact that input was typed in on a keyboard. The style of the users' spoken language probably differs from the written language. However, the main actions that the users wanted the robot to do are probably comparable to spoken interaction. In the data acquired for the following analyses, these actions were restricted by the scenario and by the tasks that the users were asked to complete.

Coding of speech

Again, the analysis below mainly focuses on what was said. Therefore, speech was annotated ignoring particularities of the speakers such as dialect. The words were annotated as they would appear in a dictionary and the sentences were annotated with a grammatically correct structure. As to segmentation, here speech encodes meaningful verbal utterances of the user directed to the robot. All utterances were segmented by the annotators into units with one communicative function (the communicative function could also be to attract the robots attention (Level 1) which is often achieved by very short utterances such as calling the robot by its name). The beginning and end of the utterances were annotated in the ELAN file[11]. However single words are not mapped exactly in time. Users' utterances like "hmm" were not annotated as they were not directed to the robot.

[11] http://www.lat-mpi.eu/tools/elan/ (27.11.2009)

The robot's utterances were retrieved from the log files. As robot speech is produced in meaningful utterances, these were taken as the basis for segmentation. Also, the robot's utterances have several functions. In general, the robot speaks to enable interaction in a natural and effective way. It tells the user what is going on within the system, shows attention, and develops a social relationship. Many utterances are task-related fulfilling the following functions:

- show that a command has been understood and the task will be carried out
- tell the user how the robot is able to complete the task and
- how the user needs to support task completion
- tell the user that the task has been accomplished

The participants were also trained to show social behaviors such as greeting and saying goodbye. In addition, how they could talk to the robot was practiced before they started to interact individually in the home tour studies. This was not the case in the object-showing studies in the laboratory. Therefore, only these will be analyzed with respect to the users' speech behaviors to determine how they tried to teach objects to the robot (see Table 3-2). The users' repertoires will be analyzed with regard to the situation, the expectations, and the tasks.

Table 3-2. Overview of analyses of speech

Object-teaching 1	no analysis
Object-teaching 2	statistical analysis of coded speech behaviors (Section 4.2.3)
Home tour 1	speech was not analyzed because users received training
Home tour 2	speech was not analyzed because users received training

3.2.2 Analysis of gesture

Speech often co-occurs with gestures. These are also crucial for interaction and shall be analyzed in the following. Based on Kendon's (2004) definition, gestures are seen here as deliberate movements with sharp onsets and offsets. They are an "excursion" in that a part of the body (usually the arm or the head) moves away from and back to a certain position. They start with preparation, continue with the prestroke hold, which is followed by the stroke. The last phase is the poststroke hold (McNeill, 2005). The movement is interpreted as an addressed utterance that conveys information. Participants in an interaction readily recognize gestures as communicative contributions. Goldin-Meadow (2003) further differentiates gestures from functional acts (for example, opening a jar). While functional acts usually have an actual effect, gestures are produced during communicative acts of speaking and do not have such effects.
If gestures are communicative acts, one could wonder why people still produce them if the other person is not present, as has been shown for telephone communication (see Argyle, 1988). The same seems to be true for the data presented here. Even though the robot cannot display gestures and the users might not be sure whether it can interpret the gestures they produce, most participants will be found to gesture (see Sections 4.2.4 and 5.1.2). This is in line with McNeill's (2005) premise that speech and gesture combine into one system in which each

modality performs its own function and the two modalities support each other or in Goldin-Meadow's words:

> *"We use speech and other aspects of the communication context to provide a framework for the gestures that the speaker produces, and we then interpret the gestures within that framework. (Goldin-Meadow, 2003, p.9)*

That gesture and speech are integrated systems can be proven because gestures occur with speech, gestures and speech are semantically coexpressive (each type of gesture has a characteristic type of speech with which is occurs), and gesture and speech are temporally synchronous (Goldin-Meadow, 2003). The integration of gesture and speech takes place quite early in the development, even before children begin to produce words together with other words.
However, many researchers have argued about this claim. According to McNeill (2005), Butterworth and Beattie (1978) are usually cited to make the point that gesture precedes speech because they found that gesture started during pauses of speech. However, they did not merely occur during pauses (McNeill, 2005). Most gestures by far occur during phonation even though more gestures occur per unit of time in pauses (the unit of time is the problem because there are probably a lot fewer pauses than speech segments). Therefore, McNeill (2005) supports the synchrony view which implies that gesture and speech are co-occurring and the stroke coincides with the most important linguistic segment. Also Argyle (1988) and Cassell (2000) argue in favor of synchrony between words and gestures and Kendon (2004) supports the synchrony view and claims that both modalities are produced under the guidance of a single plan.
While being synchronous, gesture and speech do not convey the exact same information (Goldin-Meadow, 2003). Gestures carry meaning and they are co-expressive with speech but not redundant (McNeill, 2005). Even though both modalities express the same idea they express it in different ways or in other words, co-expressive symbols are expressed at the same time in both modalities. This is due to the fact that gesture and speech convey meaning differently. Speech divides the event into semantic units that need to be combined to obtain the composite meaning. In gesture the composite meaning is presented in one symbol simultaneously (McNeill, 2005). As Goldin-Meadow (2003) puts it, speech confirms to a codified system and gesture does not. Speakers are constrained to the words and grammatical devices of a language, which sometimes fail them. In contrast, gestures make use of *visual imagery* to convey meaning. The information can be presented simultaneously, whereas in speech the presentation has to be sequentially, because language only varies along the single dimension time, whereas gesture might vary in dimensions of space, time, form, trajectory, etc. These different systems allow the speaker to create richer information. The degree of overlap (and divergence) between gesture and speech differs. It may well depend on the function of the gesture. Argyle (1975) characterizes the various roles that gesture can play in HHI (express emotion, convey attitudes, etc.), but he gives gesture no role in conveying the message itself.
Gestures do not only have a function for the listener but also for the speaker. They are a link from social action to individual cognition, can lighten cognitive load, and may even affect the

course of thought (Goldin-Meadow, 2003 and McNeill, 2005). This also explains why people gesture even if the other person cannot see the gestures.
Several categorizations of types of gestures that accompany speech have been introduced in the literature. Goldin-Meadow (2003) depicts the categorizations shown in Table 3-3.

Table 3-3. Categorizations of gesture types
(Goldin-Meadow, 2003, p.6)

Krauss, Chen, and Gottesman (2000)	**McNeill (1992)**	**Ekman and Friesen (1969)**
Lexical Gestures	Iconic Gestures	Kinetographic gestures
		Spatial movement gestures
		Pictographic gestures
	Metaphoric gestures	Ideographic gestures
Deictic gestures	Deictic gestures	Deictic gestures
Motor gestures	Beat gestures	Baton gestures

The categorizations mainly differ in the number of categories that they use (McNeill, 2005). In the following, the categorization of McNeill (1992) shall be introduced briefly. Iconic gestures represent body movements, movements of objects or people in space, and shapes of objects or people. Iconic gestures are rather concrete and transparent. The transparency depends on the speech that comes with the gesture. Metaphoric gestures present an abstract idea rather than a concrete object. Deictic gestures are used to indicate objects, people, and locations in the real world which do not necessarily have to be present. Beat gestures are beats with the rhythm of speech regardless of content. They are usually short, quick movements in the periphery of the gesture space and can signal the temporal locus in speech of something that seems important to the speaker.
However, to McNeill (2005) the search for categories seems misdirected because most gestures are multifaceted (they belong to a different degree to more than one category). Therefore, he proposes to think in dimensions rather than in categories which would additionally simplify gesture coding. Also, Goodwin (2003) stresses that one should analyze the indexical and iconic components of a gesture, rather than using categories. Even though the categorization of gesture is useful for statistical summaries, it plays a minor role in recovering gesture meaning and function. In order to determine meaning and function, the form of the gesture, its deployment in space, and the context are more important than gesture type.
Gestures can be differentiated in two basic categories: reinforcing and supplementing (Iverson, Longobardi, & Caselli, 1999). Especially deictic gestures such as pointing, references to objects, locations, and actions, can be reinforcing. The gestures label what is pointed at, but the referent can be understood with speech only. In contrast, when the gesture is supplementing the referent that is pointed at is not clear without the gesture (for example, if someone says "This" a pointing gesture is needed to clarify what is referred to with the utterance). The same differentiation has been made for iconic gestures that depict the physical characteristics of an object or action and manipulative gestures that are task-based gestures that facilitate the understanding of the action for the learner. Manipulative gestures were coded as being

reinforcing when the referent was also mentioned verbally. For Italian mothers when interacting with their children at the age of 16 and 20 months, Iverson, Longobardi, and Caselli (1999) found that very few iconic gestures were produced. Their findings replicate earlier findings of studies in the US. Altogether, mothers gestured relatively infrequently and when they did gesture, the gestures tended to co-occur with speech and were conceptually simple, i.e., less complex metaphoric gestures and beat gestures were produced. Also Rohlfing (to appear) has found that most gestures that the mothers produced in their tasks were deictic. Noncanonical relationships triggered the reinforcing function of deictic gestures. Since the robot like a child has a very restricted spatial lexicon it can be hypothesized that most gestures that the users produce are deictic. Based on this assumption, in the following the research on deictics will be introduced in more depth.

> *"Deixis refers to those linguistic features or expressions that relate utterances to the circumstances of space and time in which they occur." (Kendon, 2004, p.222)*

Thus, deictic gestures connect utterances to the physical setting. The physical world in which the conversation takes place is the topic of the conversation (Cassell, 2000). Pointing is the most obvious way to create deictic references. When pointing, the "body part carrying out the pointing is moved in a well defined path, and the dynamics of the movement are such that at least the final path of the movement is linear" (Kendon, 2004, p.199f.). Commonly post-stroke holds occur. As can be seen in Figure 3-1, Kendon (2004, p.206) identified different pointing gestures in interaction data. Index finger extended is most commonly used when a speaker singles out a particular object.

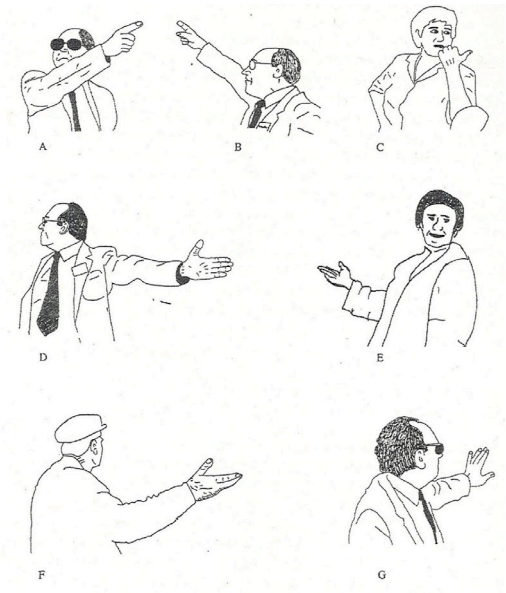


Fig. 11.4 Hand configurations used in pointing in Northant (Northamptonshire) and Campania. A. Index Finger Extended Neutral (palm vertical). B. Index Finger Extended Prone (palm down). C. Thumb. D. Open Hand Neutral (palm vertical). E. Open Hand Supine (palm up). F. Open Hand Oblique (palm oblique). G. Open Hand Prone (palm away). In Campania there seemed to be a consistent contrast in the use of the two Index Finger Extended forms which was not observed in Northant. Open Hand Oblique was not observed in Northant, Open Hand Prone (palm away) was not observed in use in Campania.

Figure 3-1. Pointing gestures according to Kendon (Kendon, 2004, p.206)

Mostly the speaker also uses a deictic word (for example, "here" or "there"). Open hand is used in their data if the object being indicated is not itself the primary focus or topic of the discourse but is linked to the topic (either as an exemplar or as a class, as the location of some activity under discussion, or it should be regarded in a certain way because this leads to the main topic). Deictic words are observed less often with open hands than when the index finger is used for pointing (Kendon, 2004). Open hand palm up is usually used when the speaker presents the object to the interlocutor as something that should be looked at or inspected. Open hand oblique, as they find in their data, is used when someone indicates an object, when a comment is being made either about the object itself or about the relationship between the interlocutors and the object (usually the object is a person in this case). Open hand prone (palm down) is used when the speaker is referring to an object in virtue of its aspects of spatial extent, or when several objects are considered together as an ensemble. Finally, pointing with the thumb occurs when the objects are either to the side or to the rear of the speaker. It is often used when a precise identification or localization of the object is not necessary (because it is known to both interlocutors or it has been referred to before). These findings show that the different pointing gestures are used in different situations. Kendon (2004) found that the gestures used in his two samples (Northhamptonshire and Campania) were not exactly similar because of the distinct situations and maybe in part because of cultural particularities.
Goodwin (2003) also describes pointing as a situated interactive activity:

> *"Pointing is not a simple act, a way of picking out things in the world that avoids the complexities of formulating a scene through language or other semiotic systems, but is instead an action that can only be successfully performed by tying the act of pointing to the construals of entities and events provided by other meaning making resources as participants work to carry out courses of collaborative action with each other." (Goodwin, 2003, p.218)*

Accordingly, pointing can only be understood with respect to the situation that must contain a minimum of two participants and pointing is at least based on the contextualization of the following semiotic resources:

"(a) a body visibly performing an act of pointing;
(b) talk that both elaborates and is elaborated by the act of pointing;
(c) the properties of the space that is the target of the point;
(d) the orientation of relevant participants toward both each other and the space that is the locus of the point; and
(e) the larger activity within which the act of pointing is embedded." (Goodwin, 2003, p.219)

As has been mentioned before, people also gesture when the other person cannot see it. However, as (a) infers, pointing only has a meaning when the other person is present. According to Clark (2003), speakers can point with any body part with which they can create a vector (finger, arm, head, torso, foot, face, eyes). With regard to (b), Goodwin (2003) has shown with a

stroke patient who could only speak three words that pointing is also possible without speech. However, the effort is much larger because speech facilitates the pointing by relating the gesture to the space in which it is occurring. Therefore, also the space (c) is an important semiotic resource. The semiotic resource (d) points to the importance of the body orientation of the participants and the objects in space. Pointing can occur together with a postural orientation. Based on the orientation, pointer and addressee form a *participation framework* (Goodwin, 2003). The framework includes orientation toward other participants and orientation toward specific phenomena located in the environment. How orientation is treated here will be elaborated on in Section 3.2.3. During the pointing activity, participants also frequently perform gaze shifts because they have to attend to multiple visual fields including each other's bodies and the region being pointed at. The pointers may initially look at the region/object they point at and then at the addressees to make sure that they respond in the correct way which would be to look at what is being pointed at. Gaze is further discussed in Section 3.2.4. Finally, (e) stresses the influence of the context which has been discussed in Section 2.1.2.

Based on the assumption that pointing activities are situated and context sensitive, it can be concluded (a) that gesture behavior has to be analyzed with the semiotic resources in mind, i.e., taking other modalities, the situation, and the context into account by assembling locally relevant multimodal packages (Goodwin, 2003); moreover, (b) that it is not advisable to use a predefined set of gestures for the analysis. Rather a specific description of gestures that occur in the data used here will have to be retrieved. Taking a first comparative look at the data from the home tour studies and from the object-teaching studies in the laboratory underlines this finding. It shall be anticipated here that the gestures used to show objects were quite different in the two settings. While in the home tour mainly pointing gestures were used, in the object-teaching study the participants preferred manipulative gestures. Moreover, the gestures used were clearly task based. While the only task in the laboratory was to teach objects to the robot, the apartment included the tasks of guiding the robot and showing rooms. Again other gestures were used in these tasks. The findings will be discussed in depth in Section 5.1.2.
Here the situatedness of gestures gives reason to introduce another differentiation that has been suggested by Clark (2003). He distinguishes pointing from placing. Like pointing, placing is used to anchor communication in the real world and it is also a communicative act. While pointing is a form of *directing* attention *to* the object with gesture and speech, *placing-for* means to put the object in the focus of attention. Clark (2003) explains placing-for with the example of the checkout counter in a store where customers position what they want to buy and where clerks expect to find the items that they need to put on the bill. These joint construals are usually established by other means than talk, i.e., items are placed in special places. Besides material things, people can also place themselves (Clark calls these self-objects and other-objects). Placing-for follows the *preparatory principle* and the *accessibility principle*.

> *"Preparatory Principle. The participants in a joint activity are to interpret acts of placement by considering them as direct preparation for the next steps in that activity." (Clark, 2003, p.260)*

> *"Accessibility Principle: All other things being equal, an object is in a better place for the next step in a joint activity when it is more accessible for the vision, audition, touch, or manipulation required in the next step." (Clark, 2003, p.261)*

Accordingly, the act of placement transmits a certain meaning (for example, I place items on the checkout counter in order to buy them) and has the function to make the object more accessible for the senses of the interlocutors. Placing-for can be divided into three phases:

> *"1. Initiation: placing an object per se.*
> *2. Maintenance: maintaining the object in place.*
> *3. Termination: replacing, removing, or abandoning the object."*
> *(Clark, 2003, p.259)*

In general, the same phases are true for actions of directing-to but the maintenance phase is very short in contrast to placing-for actions where it is continuing. Therefore, placing-for has certain advantages over directing-to: joint accessibility of signal (everyone has access to the place of the object for an extended period of time), clarity of signal (the continuing placement makes it possible to resolve uncertainties about what is being indicated), revocation of signal (placing is easier to revoke than pointing), memory aid (the continuing presence of the object is an effective memory aid), preparation for next joint action (object is in the optimal place for the next joint action). In contrast, directing to has the advantages of precision timing (some indications depend on precise timing which can more easily be done with directing-to because it is quicker), works with immovable and dispersed objects, can be used to indicate a direction and complex referents. Clark (2003) explains this with the example of pointing at a shampoo bottle saying "that company" and actually referring to Procter & Gamble. This kind of reference cannot be established with placing-for. These advantages indicate that both behaviors are valuable in different situations. The situations in the home tour and the object-teaching studies will be analyzed regarding these aspects in Sections 4.2.4 and 5.1.2 in order to determine when the participants prefer placing-for or directing-to.

Gestures in HRI

The usage of gestures has also been researched in HRI. For example, Nehaniv et al. (2005) propose five classes of gestures that occur in HRI but also mention that often gestures do not fit in one class only:

1. irrelevant (gestures that do not have a primary interactive function) and manipulative gestures (gestures that involve displacement of objects)
2. side effect of expressive behavior (motions that are part of the communication in general but do not have a specific role in the communication)
3. symbolic gestures/emblems (gestures with a culturally defined meaning)

4. interactional gestures (gestures used to regulate interaction with a partner, for example, to signal turns)
5. referencing/pointing gestures (gestures used to indicate objects or locations)

According to Nehaniv et al. (2005), it needs to be considered to whom or what the gesture is targeted (target) and who is supposed to see it (recipient).
As basis for their HRI research, Otero, Nehaniv, Syrdal, and Dautenhahn (2006) conducted an experiment in which the participants had to demonstrate how to lay the table to the robot once with speech and gesture and once using gesture only. They used the same categories as Nehaniv et al. (2005) and analyzed which gestures were used how often. However, the five categories are not specific about the movements that make up the gestures and their meaning. Hence, the applicability to the questions asked here seems limited. Moreover, manipulative gestures are from their point of view part of the category irrelevant gestures because they are not considered to have a communicative meaning. This might be true for their research questions and their scenario; however, manipulative gestures were found to be highly important by Rohlfing (to appear) who explicitly defines them as facilitating the interaction for the learner in their scenario. Therefore, they might also be important for the data analysis in the following.

Coding of gestures

The question remains of how the gestures should be coded in the data of the user studies. Goldin-Meadow (2003) points out that the gestures need to be identified in the stream of motor behavior. Approaches to describe gestures are often borrowed from sign-language literature. Accordingly, Goldin-Meadow (2003) describes the trajectory of the motion, the location of the hand relative to the body, and the orientation of the hand in relation to the motion and the body. Another scheme for coding hand orientation, hand position, and gesture phases is described by McNeill (2005, p.274f.).
Based on this information about movement one needs to attribute meaning to the gesture (which according to Goldin-Meadow (2003) is the most difficult task). To identify its meaning, the context in which the gesture occurs is important. Gesture meaning needs to be identified in relation with the task at hand. For the analysis presented below, the question arose whether the movement or the meaning of the gestures should be coded. It was decided that when the meaning was not clear right away the movement should be coded. However, for gestures with a clear meaning, a lot of time was saved in the annotation process by only coding this meaning. Therefore, for the following analyses conventionalized and unconventionalized gestures are differentiated (Kendon, 2004). Conventionalized gestures are all gestures that can be clearly associated with a meaning in a certain cultural context (for example, raising both forearms with open palms towards the approaching robot is clearly recognized as “stop” gesture). They could also be called symbols or emblems (Nehaniv et al., 2005). For these gestures it is sufficient to code the meaning. Unconventionalized gestures do not have such an unambiguous meaning attached to them. Therefore, for these gestures the movements need to be annotated in order to interpret the meaning in a second step. Also the gestures that the participants produced during the teaching tasks in the laboratory and in the apartment were coded as movements. Even

though these gestures have the unequivocal meaning of guiding the attention of the robot to the object, how this is done might make a difference in the function as Clark (2003) and Kendon (2004) have shown. To alleviate the annotation process, the movements were categorized. In a data-driven approach, typical movements were identified and categorized in both scenarios, resulting in specific coding schemes for the relevant studies (see Sections 4.2.4, 5.1.2.1, and 5.1.2.2). The coding schemes have been checked for interrater reliability.

The analysis of the gesture codings was guided by the questions of what gestures were used, i.e., what gestures did the participants' gesture repertoire consist of for certain tasks; how often were these gestures used in which situations; how long were they used; when did the participants switch between gestures; and what was the meaning of the gestures.

Table 3-4. Overview of analyses of gesture

Object-teaching 1	no analysis of gestures
Object-teaching 2	statistical analysis of gestures (Section 4.2.4)
Home tour 1	no analysis of gestures
Home tour 2	statistical analysis of gestures (pointing gestures, conventionalized and unconventionalized gestures) (Section 5.1.2)

3.2.3 Analysis of spatial behavior

As mentioned in the last section, spatial behavior and gesture are closely connected to each other and both form the participation framework. In the following, a closer look shall be taken at the spatial behavior.

Spatial behavior consists of proximity, orientation, territorial behavior, and movement in a physical setting (Argyle, 1988). All these aspects have been in the focus of the social sciences, often with respect to the influence of gender, age, and culture (for an extensive overview see Aiello, 1987; Hayduk, 1983). Many studies have shown that culture and gender influence proxemic behavior (an overview of the studies is presented in Remland, Jones, & Brinkman, 1995). With regards to age and other aspects, the studies often are not quite as conclusive, which might lead to the conclusion that other factors such as relationship between the interactants might be much more important. Moreover, the significance of spatial positions depends on the physical setting in several ways (Argyle, 1988):

- certain areas are the territory of another person or group
- certain areas are related to high or low status (for example, front row in a concert is a sign for high status)
- certain areas are associated with particular social roles (for example, the seats of people with different functions in a law court)
- rooms in a house have distinctive symbolic significance (there are rules who may enter which room)
- size, shape, and arrangement of furniture determine how close / at what angle people sit
- due to physical barriers people are much closer than they would be otherwise (for example, in the elevator)

Accordingly, spatial position depends on the physical settings but also on the situation and the task (Argyle, 1988). In interaction, spatial behaviors indicate the beginning and the end of sequences (Argyle, 1988). As Argyle (1988) found, they do not normally occur within periods of interaction. For example, a movement towards a person indicates the willingness to interact. During the interaction, spatial moves initiate the phases of the encounter (for example, someone stands up to give a speech, remains standing during the speech, and sits down again at the end of it). This finding is supported by the research of Cassell, Nakano, Bickmore, Sidner, and Rich (2001). The authors videotaped pseudo monologs (people describing their house and giving directions to a listener who only replied with minimal feedback signals) and dialogs (two subjects creating an idea for a class project). They analyzed posture shifts of upper and lower body which were motions or position shifts of the human body excluding hands and eyes. They identified posture shifts throughout both monologs and dialogs but found that they occurred more frequently at discourse segment boundaries (boundaries between topics) than within the segments and were also more energetic at the boundaries. The same was reported to be true for turn boundaries: upper body shifts were found more frequently at the beginning of turns. Moreover, posture shifts took significantly longer when finishing a topic than when the topic was continued by the other speaker. The authors further describe how the findings can be integrated in the embodied conversation agent Rea but the evaluation of the integration was not part of the paper. Nevertheless, it can be concluded that spatial behavior is an important structural feature in HHI and is therefore also taken into account here for HRI.
According to Argyle (1988), the choices that we make regarding spatial behavior are due to the organs that we use for sending and perceiving certain signals. For example, people giving a speech will stand up so that the listeners can hear them well and also see them. However, on top of varying the spatial behavior to best overcome perceptual limitations, there is still some room for variation that can be exploited to express feelings towards another person or the situation in general. Despite all speakers standing up for their speeches in a plenum, there are significant differences in their body postures and of course, also in gazing behavior, speech, and gestures that signal how the person feels about the situation. Therefore, it can be concluded that movements in space are used as moves in social interaction that have a communicative function. This function depends on the conscious recognition by the participants:

> *"Spacing behaviors, then, to be communicated must be consciously recognized and viewed as a message by at least one of the participants." (Burgoon & Jones, 1980, p.198)*

Accordingly, spacing behaviors must be consciously received as messages (Burgoon & Jones, 1980). The messages people send with their spatial behavior were the basis for the *theory of personal space expectations and their violations*. Propositions of the theory were the following (Burgoon & Jones, 1980):

- "Expected distancing in a given context is a function of (1) the social norm and (2) the known idiosyncratic spacing patterns of the initiator." (p.199)

- "The effects of violations of expectations are a function of (1) the amount of deviation, (2) the reward-punishment power of the initiator, and (3) the threat threshold of the reactant." (p.202)

As can be seen from the propositions, Burgoon and Jones (1980) suggest expectations towards social norms and the individual person as the basis for appropriate behavior. Expectations reflect our predictions about how others will behave. They serve as a level of adaptation and comparison for a given context.

> *"Factors such as (1) the participant's sex, age, race, status, and degree of acquaintance, (2) the situation's degree of formality, topics, and task requirements, and (3) environmental constraints all play a role in defining what is normative distance for a given conversation." (Burgoon, 1983, p.78)*

Thus, the participants' characteristics and the situational and environmental constraints are the basis for the normative behavior and the expectations related to it. If these expectations are violated, the reaction towards the violation depends on the degree of the deviation between expected behavior and performed behavior, the power of the person who exhibits the behavior to reward and punish the perceiver, and the individual characteristics of the person who is exposed to the behavior (see Section 2.2.4). Reactions to violations may range from highly positive to highly negative and, thus, have positive or negative consequences. For example, if the head of the department comes close or even touches the employee this might be a really positive violation because closeness signals liking and appreciation. In contrast, if the head of the department always takes a position further away than expected, this signals a violation of the expectation that has a negative valence. The example shows that the violation also depends on its direction (too close, too far). However, as Burgoon (1983) additionally points out, what is perceived as being too close or too far is very individual. The individual expectations are assumed to be the result of each person's experience with normative behaviors and they are also influenced by the knowledge about the other person's expectations.
As long as the expectations are realized, proxemic behavior is rather subconscious. However, as soon as the expectations are violated "the person will become aroused physiologically and psychologically and attempt to restore homeostasis or balance" (Burgoon, 1983, p.79). Most often this will be attempted by a movement in space. It can be assumed that the same is true for HRI. However, in HRI the violation of the expectations might additionally lead the participants to question the abilities of the robot. Assuming that they do not believe that the robot violates their expectations willingly, they will probably question the abilities of the robot, asking whether it sees them, has understood their command, is able to stop, or to maintain a safe security distance.

Spatial behavior in HRI

Some of the questions just mentioned have already been addressed in HRI research. Spatial distances were one main research activity in the EU project Cogniron that focused on

"observing a correct attitude towards humans, especially with respect to interaction distances and personal spaces" (www.cogniron.org, 24.09.2009). In this context, a comfort level device was developed to measure how comfortable the user felt with the robot keeping a certain distance or approaching from different directions (see, for example, Koay, Walters, & Dautenhahn, 2005; Syrdal, Koay, Walters, & Dautenhahn, 2007; Walters, Koay, Woods, Syrdal, & Dautenhahn, 2007).

In general, it was concluded that the participants prefer Hall's personal distance (0.46 to 1.22m) when interacting with a robot (Hall, 1966; Hüttenrauch, Severinson-Eklundh, Green, & Topp, 2006). Thus, their spatial behavior complies with HHI. However, the picture becomes a little more complicated when researching more complex situations. Koay, Syrdal, Walters, and Dautenhahn (2007) found that the users felt discomfort when the robot was closer than three meters while they were performing a task. This distance is within the social zone reserved for human-human face-to-face conversation. The result points to the connection between behavior and situation. It was confirmed by Walters et al. (2007) who showed that different social approach rules apply depending on whether the interacting human is sitting, standing in the open, or against a wall or obstacle. Dautenhahn et al. (2006) have presented two studies that investigated how a robot should best approach and place itself relative to a seated human subject in the scenario of the robot fetching an object that the human had requested. The studies indicated that most subjects disliked a frontal approach and preferred to be approached from either the left or right side with a small overall preference for a right approach. However, the expectations of the users were also found to be influenced by the tasks that they had to perform. Syrdal et al. (2007) report a study with a robot approaching a person in three conditions: verbal interaction, physical interaction, and no interaction. The approach direction again varied. They found that people tolerated closer approaches for physical interactions (for example, performing a joint task or handing over an object), as opposed to purely verbal interactions (when giving commands, etc.) or where the robot is just passing by and not directly interacting with the human. Moreover, tendencies towards gender and personality differences were reported. These findings are in line with Burgoon's (1983) theory (see Section 2.2.4).

The work presented so far focused on approaching directions and proximity. Also body orientation has been of interest in HRI research. Suzuki, Morishima, Nakamura, Tsukidate, and Takeda (2007) evaluated the influence of agent body orientation on task completion. However, their work is restricted by the fact that they only distinguished two orientations (facing the user/ back towards user). Hüttenrauch et al. (2006) discuss Kendon's F-Formation system (Kendon, 1990) which is much more natural in interaction (see Figure 3-2).

It is based on the observation that people usually group themselves in certain formations when interacting. F-formation describes the kind of formation when people form a shared space between them which they can all access equally well. Hüttenrauch et al. (2006) analyzed their data, which was also acquired within the home tour scenario, for Kendon's F-Formation arrangements. To begin with, they found what they called "micro-spatial adaptations" before many new interaction tasks which is in line with the assumption discussed above that many adaptations of spatial behavior occur at discourse segment boundaries. Table 3-5 shows the re-

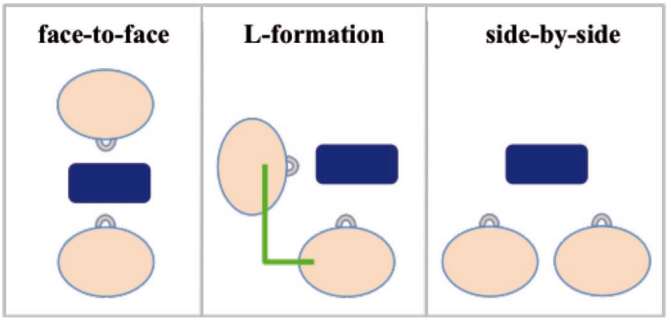


Figure 3-2. Kendon's F-formation
(Kendon, 1990)

Table 3-5. Results of analysis of body orientation with Kendon's F-Formation
(Hüttenrauch et al., 2006)

task	vis-à-vis	L-shape	side-by-side
follow	82	13	4
show	71	45	0
validate	58	27	1

sults for all tasks (the robot following the human, the human showing objects/places to the robot, the human validating that the robot had learned the object/place).As can be seen in Table 3-5, the vis-à-vis formation was dominant in all tasks. The L-formation only occurred to a considerable degree in the showing task. However, it should be noted that this number includes objects and places. It can be questioned whether the behavior in both cases is similar. This question will be discussed in depth in Section 5.1.3. As the authors themselves note, more formations occurred that could not be coded with this scheme. That is why a more flexible coding scheme will be introduced here.

Coding of spatial behavior

Since Kendon's F-Formation only allows for the coding of three different orientations, a more flexible approach was sought for the following analysis. One such approach is the sociofugal-sociopetal (SFP) orientation as Hall (1963) calls the orientation that separates or combines people, respectively (see Figure 3-3).

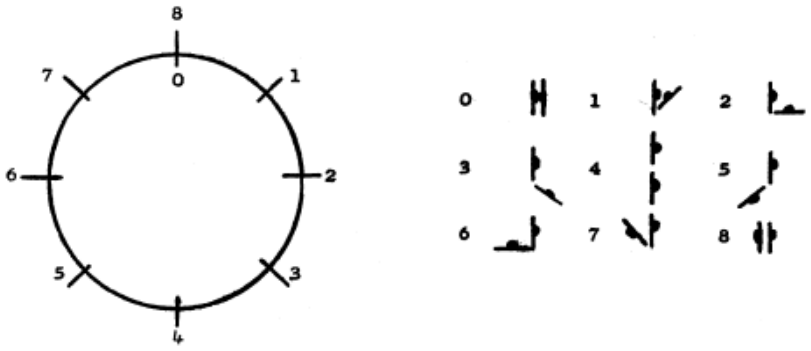


FIG. 1. SFP axis notation code.

Figure 3-3. Hall's SFP axis notation code
(Hall, 1963, p.1009)

Orientation

> *"is the angle one person faces another, usually the angle between the line joining him to another and a line perpendicular to the plane of his shoulders, so that facing directly is 0 degrees. (This refers to the orientation of the body, not the head or eyes.)" (Argyle, 1988, p.170)*

In theory, an endless number of orientations are given if intervals are specified on a circle. However, as Hall (1963) notes, "the observer is interested in recording only those distinctions which are operationally relevant to the participant" (p. 1008). Hall (1963) and colleagues experimented with a number of overly sensitive systems before they decided on an 8-point compass face (see Figure 3-3). 0 represents two subjects face-to-face, 8 back-to-back. Shuter (1977) modified Hall's notation system to a scale of 0 to 12 corresponding to the hours on a clock. Each score also represented a different shoulder orientation from face-to-face to back-to-back.

For the research presented here, eight intervals seem to be operationally relevant. The SFP axis notation code still needs to be modified because it does not allow to determine which interactor stands on which side. While this differentiation might play a minor role in HHI, it is assumed to be important for HRI because of the asymmetry of the interaction and the different behaviors that are performed by the human and the robot (that differences between spatial behaviors actually exist will be shown in Section 5.1.3). Moreover, it was found to be necessary to consider the facing direction of both the human and the robot for two reasons. From the point of view of the robot, the facing direction is crucial because the robot loses the users as soon as it cannot perceive them anymore and the users need to repair this problem before the interaction can continue. Even though the robot is not (yet) programmed to adapt its body orientation to certain tasks based on a social model of the interaction, the teaching of locations requires it to turn around to improve its representation of the environment. From the point of view of the user, the facing direction is important because it indicates the users' expectations about the robot's abilities and their trust in what the system is doing. The users will most probably closely monitor the robot not turning the back towards it if they have doubts about its ability to follow without doing any harm.

Based on these assumptions, for the following analysis a coding scheme with 23 different codes was used that not only allows to code body orientation as such, but also allows to identify who has changed the orientation (human, robot, or both) and which direction each interaction partner is facing (see Appendix A). To capture the dynamics of the body orientation in depth, static orientation and dynamic changes of orientation of both the user and the robot were distinguished. This allowed to code changes of body orientation of the robot and the human that happened at the same time, and changes of body orientations that happened one after the other.

In addition, also the orientation of the user towards the objects was coded for the object-teaching task in the apartment (it was not coded in the laboratory because there the setting was static) in order to determine exactly how the users arranged the participation framework and whether there were differences with respect to the objects that had to be shown. For the annotation of the object position in relation to the user, the eight orientations specified by the

SFP axis notation code were considered to be adequate. The orientation towards the objects was also coded with regard to the orientation of the shoulders of the user.

Based on these annotations, the body orientation of human and robot could be determined at all times of the interaction. It is of interest here what orientation the users chose, whether the orientation depends on certain tasks, and what the switches in orientation tell us about the importance for spacing in the interaction. With respect to the last point, the question of how much the participants try to reach a certain position needs to be considered (how much time do they spend on spatial adaptation, how often do they change their orientation, and what does this imply for their need to attain the robot's attention?). These questions are analyzed in the following without taking proxemics (the distance between the user and the robot) into account. There surely is a correlation between both measures, for example, the closer a person stands to another person, the more indirect the body orientation usually is. However, there are two main reasons to focus on body orientation here: as has been mentioned before, much work has already been conducted on proxemics in HRI while body orientation has not played such an important role. Yet, body orientation is considered here as an important factor because it structures the interaction and, hence, implies a great deal of information that could be useful for the robot to better understand what the user wants to do. Moreover, the robot is not yet able to adapt its requirements concerning the human body model to the situation, i.e., in all situations it needs the same percepts to represent the user. However, it can be assumed that body orientation differs between tasks. That is why research into body orientation can help to develop a more flexible human body model. On the other hand, it can also help to improve the robot's behavior model and to enable it to adapt its behavior to the situation and, hence, to improve the interaction. Thus, the robot also needs knowledge about the social function of spatial behavior.

To analyze the data with these aspects in mind, the body orientation was annotated in the data of the second iteration of the home tour study in the apartment. To test the reliability of the coding scheme, interrater agreement was calculated (see Section 5.1.3).

Table 3-6. Overview of analyses of body orientation

Object-teaching 1	no analysis of body orientation
Object-teaching 2	no analysis of body orientation
Home tour 1	no analysis of body orientation
Home tour 2	statistical analysis of body orientation (Section 5.1.3)

3.2.4 Analysis of gaze

Gaze is of central importance in social behavior (Argyle, 1988). According to Argyle (1969), the direction of gaze indicates the direction of attention. Therefore, where a person looks is crucial. Moreover, tasks like object-teaching require joint attention towards the object. Nagai, Asada, and Hosoda (2006) define joint attention "as looking at an object that someone else is looking at by following the direction of his or her gaze" (p.1166f.). Children acquire this ability at about 18 months of age. It is understood here as a cue that allows to determine visually that interactants are talking about the same thing.

Gaze differs from other modalities because it is a non-verbal signal and at the same time a way of perceiving others, their faces, and expressions. Gaze is both a signal and a channel. People use the direction of their gaze to designate persons or objects that they are attending to. Hence, gaze is one kind of directing-to action (Clark, 2003, see Section 3.2.2). However, it is not effective unless it is registered by the other interactor. Therefore, it is usually grounded by mutual gaze. Goodwin's (1981) conversational rule states that the speaker has to obtain the gaze of a listener to produce a coherent sentence. To attract the gaze, the speaker can either make a restart at the beginning by uttering a phrase before uttering the coherent phrase or by pausing. Moreover, gaze is often used along with face direction, torso direction, and pointing (Clark, 2003).
Even though gaze is an important signal of attention, speakers look intermittently at their listeners (Argyle, 1988). Reasons for frequent gaze shifts are that looking at the listener all the time could lead to cognitive overload or the arousal caused by the gaze could infer with the cognitive planning. Argyle (1988) provides basic statistics of people's gaze in emotionally neutral conversations with a distance of about two meters (see Table 3-7).

Table 3-7. Statistics of gaze behavior (Argyle, 1988)

individual gaze	60% of the time
while listening	75% of the time
while talking	40% of the time
length of glance	3 seconds
mutual glance	30% of the time
length of mutual glance	1.5 seconds

The table shows that the amount of gazes varies (Argyle, 1988). For example, the interactors look less at each other when there are other things to look at, especially when there is an object of joint attention. This indicates that in an object-teaching task, where the object is in the focus of attention, the users will look less at the robot (see Sections 4.2.5 and 5.1.4). Moreover, the relationship between the interactors influences the amount of gaze. Argyle (1988) found that strangers who were two meters apart looked at each other 40% of the time. He also reported that this number might be higher for couples and it is also higher if the individuals like each other in general. Also, the spatial situation is of importance; greater distances lead to more gazes. Finally, the personalities of the interactors determine the amount of looking. Dominant individuals were found to look more while they talk (Argyle, 1988).
Next to these situational constraints, also the conversation itself structures gaze behavior. The main reason that people gaze is to obtain additional information especially about what was just said. That is why more gaze is necessary at the end of the utterance because then feedback is needed. Hence, speakers look more at the end of utterances and look away at the beginning of utterances, especially if they have been asked questions. In turn, listeners typically look 70% to 75% of the time in quite long glances of about seven to eight seconds, because they try to pick up the non-verbal signals of the speaker. However, they also look less than 100% of the time to decrease cognitive load and arousal (Argyle, 1988).

At the end of each turn, often (62%) a *terminal gaze* occurs (Kendon, 1967). A terminal gaze is a prolonged gaze at another just before the end of long utterances. If this gaze is not exchanged, the transition to the next speaker takes longer.
According to Argyle (1988), the usual measure of gaze is *mutual gaze,* i.e., the percentage of time that is spent looking at another in the area of the face. Moreover, one can measure looking rates while talking and while listening, average length of glances, pattern of fixation, pupil dilation, eye expression, direction of gaze-breaking, and blink rate. Usually the measures are chosen according to the research question at hand.
Brand, Shallcross, Sabatos, and Massie (2007) use eye gaze as a measure of interactiveness in child-directed demonstration. They count gaze bouts (gaze shifts from elsewhere to the face of the partner) per minute, measure gaze duration (the percentage of the demonstration spent gazing at the partner), and compute average gaze length (average length of each gaze bout). Based on this work, Vollmer et al. (2009) evaluated eye gaze in studies with the virtual agent Babyface as a measure of contingency. The results were then compared to adult-child and adult-adult interaction. Comparable to the data presented here, their study was also based on a teaching scenario. However, the participants did not teach objects but actions. For their data, Vollmer et al. (2009) computed the frequency of eye-gaze bouts to the interaction partner and the object (eye-gaze bouts per minute), the average length of eye-gaze bouts to the interaction partner and object, and the total length of eye-gaze bouts to the interaction partner and object (percentage of time spent gazing at agent or object).

Gaze in HRI

Some work has also been conducted on gaze in HRI. Staudte and Crocker (2008) carried out an experiment focusing on the production of robot gaze in a scenario where the robot pointed out objects to the participants either correctly or falsely. They measured the reaction time of the human depending on the gaze direction and the speech of the robot and showed that in the case that the robot's gaze or speech behaviors were inappropriate, the interaction slowed down measurably in response time and fixation distribution.
Sidner, Kidd, Lee, and Lesh (2004) report a study in which the robot Mel gazed at the participants. They coded shared looking (mutual gaze and both look at the same object) and found a positive relationship between shared looking and engagement of the interactors. Moreover, they showed that looking and gestures are more powerful than just speech to achieve engagement. They get people to pay more attention to the robot and they may also cause people to adjust their gaze behavior based on the robot's gaze.
Green (2009) reports gaze as one mean to establish contact in a home tour scenario. Users "look for" feedback in order to find out whether a command has been received correctly by the robot. Therefore, gazing at the robot indicates that some kind of feedback is necessary. Green (2009) concludes that the system should provide continuous information on its status of availability for communication. This status can be provided with the gaze of a camera because just like a human eye the camera behavior of the robot can be used to grab the user's attention and to make contact. This assumption is also followed in the work with BIRON.

To conclude, gaze has been shown to have several functions in HRI such as communicating status, engaging the user, and making interaction more efficient. These functions have been evaluated with different measures.

Coding of gaze

The general question regarding gaze in the following analysis is how the situation and the expectations influence human gazing behavior. Therefore, it will be evaluated where the users gazed in general and in different phases of the interaction. Moreover, it will be determined how gaze related to other modalities (speech, gesture) since strong interrelations between all modalities can be expected. The users' gaze behavior was annotated using three distinct directions: gaze at the robot, gaze at the object of interest, and gaze somewhere else. Since only three gaze directions were differentiated and many studies have found interrater agreement of over 90% for coded gazing behaviors with more diversified schemes (Argyle, 1988), interrater reliability has not been calculated.

Table 3-8. Overview of analyses of gaze

Object-teaching 1	no analysis of gaze behavior
Object-teaching 2	statistical analysis of gaze direction (Section 4.2.5)
Home tour 1	no analysis of gaze behavior
Home tour 2	statistical analysis of gaze direction (Section 5.1.4)

3.2.5 Analysis of integrated modalities and interaction structure

Next to the analysis of single modalities, also their interplay will be analyzed in order to identify how the users typically combine behaviors of different modalities to form consistent multimodal behaviors. As has been highlighted in the context of the dynamic concept of situation, modalities have frequently been found to alter each other (see Section 3.1.1.2). If the robot acquired knowledge about the interplay of modalities, it could more easily identify the intention of the user.

Next to their interplay, the modalities have also been analyzed for different phases (present, wait, answer, react) and courses of the interaction (positive/negative) in the second object-teaching study and for different tasks in the second home tour study. The results will show whether these situational differences influenced the behavior of the participants (see Sections 4.2.2, 5.1.1).

Finally, the approach allows for comparison of the data of different studies. In the following, the gaze behavior in the second object-teaching study will be compared with the gaze behavior during the object-teaching task in the second iteration of the home tour study (see Section 5.1.4.4). Since both studies have been annotated with the same criteria and evaluated with the same measures, a comparison can easily be done without having to prepare the data for it in an extra step. Also the gesture behaviors will be compared for the second object-teaching study and the object-teaching task of the second iteration of the home tour (see Section 5.1.2.3). As two different coding schemes were used, the comparison is based on the types of gestures and not on single behaviors. It will show whether the two situations require different types of gestures.

Table 3-9. Overview of analyses of the interplay of modalities and interaction structure

Object-teaching 1	no integrated analysis of modalities
Object-teaching 2	comparison of speech, gesture, and gaze between positive and negative trials and in the phases present, wait, answer, react (Sections 4.2.3, 4.2.4, and 4.2.5) analysis of interrelation between gesture and speech, gaze and speech, gesture and gaze (Section 4.2.6)
Home tour 1	no integrated analysis of modalities
Home tour 2	comparison of gesture, body orientation, and gaze behavior between the tasks (Sections 5.1.1, 5.1.2, 5.1.3, and 5.1.4) interrelation between gaze and body orientation (Section 5.1.4)
Object-teaching 2 and Home tour 2	comparison of gestures (Section 5.1.2.3) comparison of gaze behavior (Section 5.1.4.4)

3.2.6 Statistical AnaLysis of Elan files in Matlab (SALEM)

The goal of analyzing the modalities and their interplay in such depth as introduced in the previous sections required developing an approach that allows to conduct all analyses in an efficient way. Since ELAN has been used as a tool to represent videos, automatic log files, and manual annotations of the user studies in a synchronized way it should be included in the analysis process. ELAN represents the annotations as occurrences in time with a concrete beginning and ending. This is advantageous compared to the analysis of discrete micro-behaviors that are coded in certain time steps (for example, second-by-second) as reported by Dautenhahn and Werry (2002), because when merely analyzing "data points" much information is lost and the results are approximated values. When the intervals are smaller, the approximation gets better; however, the coding is then more laborious. That is why continuous coding is preferred here.

ELAN is basically a tool for annotation and not for quantitative analysis. Now, to analyze the data, one can go through the files manually or, alternatively, ELAN offers some formats to export the data all of which are text-based. This means that a text file has to be exported to later import it in another software package (for example, Matlab, SPSS) for analysis. This approach is rather laborious because whenever a change is made to the ELAN file, it has to be exported and imported to the analysis software again. Moreover, when exporting the text file, it is not possible to see right away whether all annotations are correct or if something was mistyped or annotations were made that do not comply with the coding scheme. To work around these problems, the SALEM toolbox was developed to parse ELAN files directly and to conduct statistical analyses with Matlab. Thus, it closes the cycle of importing all automatic log files into ELAN, importing the videos into ELAN, annotating manually in ELAN and analyzing the data in an efficient way (see Figure 3-4). In this process, one main advantage of the SALEM toolbox is that it allows comparing annotations of different modalities, structural features of the interaction, or whatever has been annotated. For example, the video can be used to manually annotate human speech which can then be compared to the speech that the robot understood because both are represented in one file that can be analyzed in Matlab. Analyzing the data with

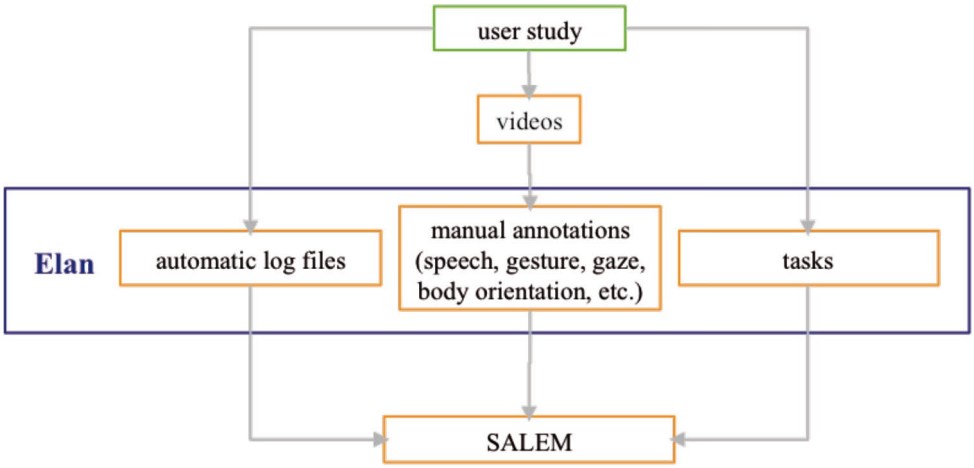


Figure 3-4. SALEM (Statistical AnaLysis of Elan files in Matlab) in the analysis process

the SALEM toolbox not only alleviates the analysis process but also increases the consistency of the analysis because all evaluations are based on the same statistical grounds.
Mainly based on requirements that arose during the analysis of the data presented here, SALEM to date has the following functionalities:

Parsing, displaying structure of parsed files, and slicing:

- parsing single ELAN files or a set of ELAN files at once (which allows for the analysis of the data of single users, groups of users, groups of trials that belong to certain conditions, and all users of an experiment or a study)
- display the names of all tiers in the parsed files
- plot the annotations in the tiers
- slice the files with respect to time (specifying one or more beginnings and endings of timeslots)
- slice all annotations of a single tier (for example, if the file is sliced on the basis of the tier human speech, then in all other tiers only the annotations that overlap with instances of human speech are taken into account)
- slice the files with respect to one or more values of the annotations in a single tier (for example, slice all annotations of gaze that have the value "1" which means that the user is looking at the robot)
- examine one specific annotation in a tier (for example, the 12th annotation in the tier gaze direction)

Analyzing:

- descriptive statistics for all data of the parsed files or the slices (for each tier):
    - count of annotations (number of occurrences)
    - minimum duration of the annotations (in seconds)
    - maximum duration of the annotations (in seconds)
    - mean duration of all annotations (in seconds)
    - median of the durations (in seconds)
    - overall duration of all annotations (in seconds)

- variance and standard deviation of the duration of all annotations
- beginning of first annotation (in seconds)
- end of last annotation (in seconds)
- overall duration of all annotations as a percentage of the time between the beginning of the first annotation and the end of the last annotation

- the descriptive statistics for slices additionally include for all tiers
  - count and percentage of time that the annotations in a tier overlap with the reference tier for four types of overlap: the annotation extends the annotation in the reference tier (begins before the annotation and ends after the annotation in the reference tier); the annotation is included in the annotation in the reference tier (begins after the annotation in the reference tier and ends before the annotation in the reference tier); the annotation begins before the annotation in the reference tier begins and ends before it ends; the annotation begins after the begin of the annotation in the reference tier and ends after the end of the annotation in the reference tier (see Figure 3-5)

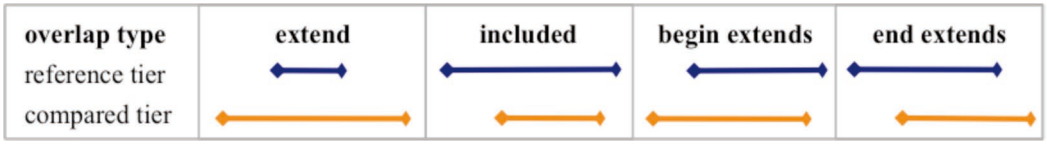


Figure 3-5. Overlap types of annotations in different tiers

- statistics for the annotated values in a certain tier:
  - duration of all annotations for all values
  - descriptive statistics for all values (see descriptive statistics for all tiers)
  - T-tests for the duration of the annotations and the duration of the overlap
  - predecessor transition matrix for all values in the tier (percentages with which all values preced all other values) (see Appendix B)
  - successor transition matrix for all values in the tier (percentages with which all values succeed all other values) (see Appendix B)

3.3 Analysis of the tasks of the interaction with Systemic Interaction Analysis (SInA)

While SALEM leads to many insights regarding the modalities, it cannot be used to analyze the situation in its overall complexity and to develop an all-embracing picture of the interaction and the problems that occur. That is why the SInA (Systemic Interaction Analysis) approach was developed to analyze the interaction in a systematic and qualitative way. More specifically, it allows to compare situations on the interaction level which describes the interplay between the system and its user in the context of a certain task; but also includes knowledge about what happens within the robot, i.e., on the system level. The system level depicts the robot components and their interplay. The components need to be evaluated in the system context because of the complexity that arises based on the interplay of processes in a system and the perceptual limits caused by the system sensors. A component that works well when tested

independently might turn out to be problematic once integrated in a system with other components. Moreover, the underlying model of the component might appear adequate from a technical point of view but might, however, not be appropriate for the interaction with a user under certain conditions. Therefore, a systemic analysis approach is required to address system and interaction level of HRI at the same time.

In this context, SInA is used to identify the task structure, i.e., what the robot does, what happens within the robot, and what the users do. The method allows for a careful description of the interaction and, thus, helps to determine the relations between the user's and the system's behavior. Deviations from the prototypical task structure based on inappropriate expectations of the user and inadequate design of the robot can be identified. In the following, the theoretical background for SInA is introduced and the SInA evaluation process is described.

3.3.1 Theoretical background of SInA

The SInA approach shares many characteristics with traditional interaction analysis used in HHI. The interaction analysis approach is interdisciplinary and it is applied to the empirical investigation of interaction between interlocutors with each other and their environment (Jordan & Henderson, 1995). The predominant research interest of interaction analysis, to investigate how people make sense of each other's actions and how this can be seen in their actions, is also shared here (Jordan & Henderson, 1995). However, the interactions are influenced significantly by the fact that one interactor is a robot.

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 annotations. The IAP consists of six layers of information such as annotations 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 SInA. However, the IAP tool does not take the system level into account. Therefore, 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. In the following, the units of analysis are tasks identified with the help of task analysis (TA) whereas Burghart and colleagues break an interaction down into "phases" (opening, negotiation, and ending of the interaction).

TA methods were chosen as the main background for SInA, because according to Steinfeld et al. (2006) HRI is task-driven. TA has been used in market research, product testing, and also in human-computer interaction for a long time.

> *"A task is regarded as a problem to be solved or a challenge to be met. A task can be regarded as a set of things including, a system's goal to be met, a set of resources to be used, and a set of constraints to be observed in using resources." (Shepherd, 2001, p.22)*

Accordingly, a task is a goal that has to be achieved with the help of a system. Tasks can be analyzed by means of TA.

> *"Task analysis can be defined as the study of what an operator (or team of operators) is required to do in terms of actions and/or cognitive processes, to achieve a system goal." (Kirwan & Ainsworth, 1992, p.1)*

In other words, TA provides a picture of the system from the perspective of the user, which can be used to improve task completion and to integrate the human into the system design. TA helps to generate hypotheses about sources of inadequate performance of system and operator (Shepherd, 2001). According to Hackos and Redish (1998), user and task analysis focus on understanding how users perform their tasks. Several questions need to be answered: What are the users' goals and what do they do to achieve them?; What are the personal, social, and cultural characteristics of the users?; What previous knowledge and experiences do they have?; How are they influenced by their physical environment?
This view is strongly user centered and, thus, aims to improve the user experience. But how can changes be made to the systems that lead to higher user satisfaction? To approach this question Shepherd (2001) has proposed a hypothesis and test cycle of task analysis (see Figure 3-6).

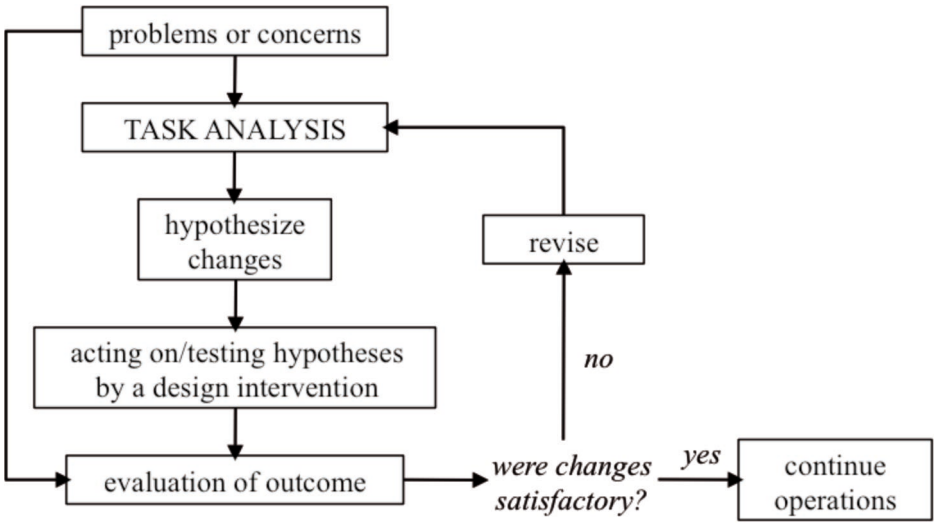


Figure 3-6. Task analysis cycle
(Shepherd, 2001, p.19)

This cycle depicts the process of improving the system iteratively by making changes, evaluating them, and – if necessary – making more changes. This approach is valid for all kinds of TA. Two subtypes of TA are hierarchical task analysis (HTA) and cognitive task analysis (CTA). HTA is characterized by its generic nature and applicability to any type of task in any domain. In HTA, tasks are essentially defined by *goals* that a user wants to achieve rather than by actions. Complex tasks may be analyzed by decomposing them into subgoals (Annett & Stanton, 2000) or subtasks (Kirwan & Ainsworth, 1992). According to Kirwan and Ainsworth (1992), HTA produces a hierarchy of operations (different things people must do within a system) and plans (statements of the necessary conditions to undertake the operations). Plans in

HTA coordinate how subordinate operations are organized in order to meet their common goal. They comprise timing and sequencing relationships (Shepherd, 2001, p.42). HTA also helps to understand how the task context affects behavior. Context influences are: goal context; frequency, predictability and coincidence; priming and sharing information; and decision outcome (Shepherd, 2001, p.77).
While HTA focuses on the subtasks or subgoals, cognitive task analysis (CTA) aims to understand how human cognition makes it possible to get things done, or in other words:

> *"Cognitive Task Analysis studies try to capture what people are thinking about, what they are paying attention to, the strategies they are using to make decisions or detect problems, what they are trying to accomplish, and what they know about the way a process works." (Crandall, Klein, & Hoffman, 2006, p.9)*

Accordingly, CTA focuses on human thinking, attention, decision-making, and knowledge. CTA methods are used to determine what people know and how they know it. Thus, CTA focuses on single users and how they solve tasks. In contrast, HTA analyzes the task structure. Therefore, CTA may explain human behavior that HTA may not cover. On the other hand, CTA might focus on the human and neglect the context (Shepherd 2001). Since the context is much broader than the processes within specific users, CTA might not lead to results that apply to all people. Hence, the methods with their advantages and shortcomings can and should be combined. This is also done here. In the following, the term TA refers to traditional TA as well as to thoughts inspired by HTA and CTA.
In recent years, TA has been introduced to HRI; for example, in the context of situational awareness and user roles (Adams, 2005). Severinson-Eklundh, Green, Hüttenrauch, Oestreicher, and Norman (2003) 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. 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. However, TA is also useful in these respects. In particular, HTA as developed by Annett and colleagues (see Stanton, 2006) is a useful tool for the analyses presented in the following because it can be used to define tasks and subtasks and their interrelationships.

SInA can also be used in conjunction with other methods such as ethnomethodological Conversation Analysis (CA) (Lohse, Hanheide, Pitsch, Rohlfing, & Sagerer, 2009). 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. In its traditional form, CA serves to analyze the processes and the structure of an interaction. More recently, a central focus has been on 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).

3.3.2 SInA evaluation process

The goal of SInA is to obtain a detailed description of system and interaction level in order to incorporate the findings in the system design and to improve HRI with autonomous systems. The focus of the evaluation is mainly to understand and assess (a) how the system performs in collaboration with users and (b) why it performs in the observed way. In this process, concrete questions need to be answered:

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

These questions can be addressed with the SInA cycle (see Figure 3-7) that is explained in the following.

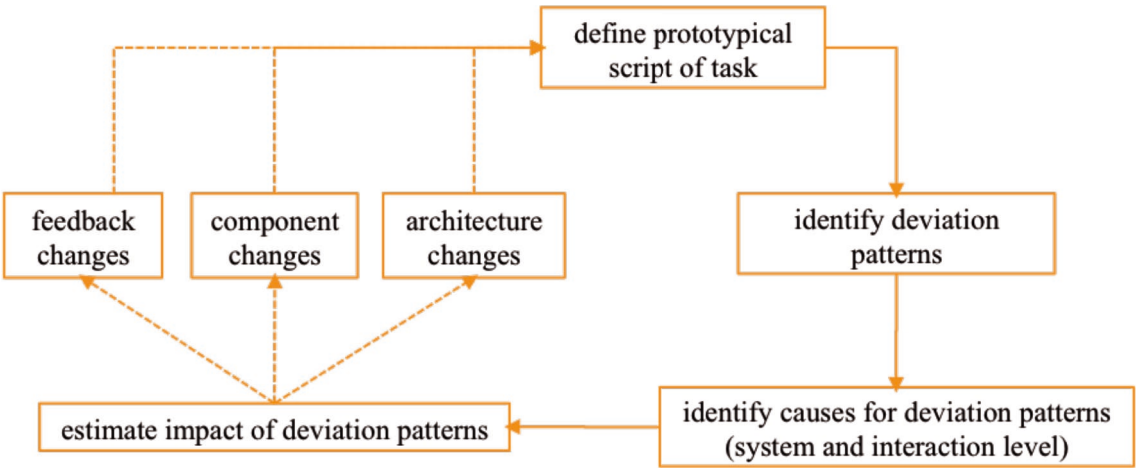


Figure 3-7. Systemic Interaction Analysis (SInA) cycle

Prototypical interaction script

TA is applied to identify all tasks and subtasks. A task is selected and all its instances are analyzed 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. Similar to 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. In the ideal case, the prototypical interaction script is in accordance with the prototype that the users have in mind (see Section 2.2.3). If it is not, the users' expectations might lead to deviation patterns.

Deviation patterns

In the second step of SInA, cases in which the interaction deviates from the prototypical script are identified. Deviations are to be expected if a component is tested in an integrated system with real users and its model needs to be adapted. There are several possible reasons for the occurrence of deviation patterns:

- the robot perceives the situation wrong and its behavior is inappropriate
- the robot cannot interpret the user behavior at all
- the robot gives feedback that triggers inappropriate expectations in the user
- the robot does not provide the user with enough feedback and the user does not know what is going on
- the user has wrong expectations that are not related to the robot's feedback
- the robot perceives the situation correctly but is not able to resolve it on its own

Deviating cases are observed on the interaction level, and their causes are traced back to the system level where the responsible components are identified. This constitutes the core idea of SInA. In order to verify that deviations have not occurred by coincidence, further examples of each phenomenon need to be identified. Deviations that occur only once are not included in the SInA procedure. However, they can be noted for later analysis within other approaches.

In the next step, groups of deviating cases are defined that are called 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. A categorization of the patterns can be derived by clustering them according to the robot's functions (speech understanding, person perception, etc.).

Within this second step, quantitative measures of the occurrence frequencies of the patterns are obtained 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. The 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.

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 the users' expectations and behaviors 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 3-8).

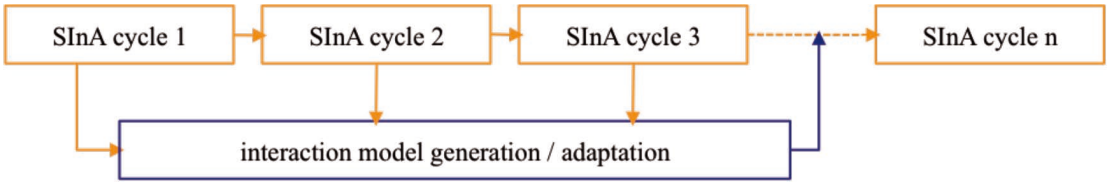


Figure 3-8. Short-term and long-term effects of SInA

The changes mainly depend on the question of why the deviation pattern occurred. For example, if the robot perceived the situation wrongly, its model of the environment and the user's behavior might have to be adjusted; if the user has wrong expectations, the robot feedback needs to be adjusted in order to correct the expectations. Thus, the deviation patterns help to identify what changes need to be made. Finally, the changes need to be evaluated. This step can only be achieved by reiterating the whole SInA procedure. Therefore, as Figure 3-7 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.

In the following, the SInA procedure is applied to the data of the home tour studies where the robot had a lot of abilities and operated autonomously. Since the system level analysis plays a major role in the SInA procedure, it cannot be applied to Wizard of Oz studies such as the second object-teaching study.

Table 3-10. Overview of Systemic Interaction Analyses

Object-teaching 1	no SInA
Object-teaching 2	no SInA
Home tour 1 and 2	conjunct SInA (Sections 5.2 and 5.3)

3.4 Visualizing interaction sequences for analysis

The methods introduced so far focus on the quantitative and structural analysis of HRI. However, sometimes they do not sufficiently display differences in sequences that include the same behaviors of the repertoires. In other words, the same behaviors occur but they are timed differently. In these cases, one solution is to compare the sequences on a time scale. ELAN files are an appropriate input for such visualizations because they capture the different modalities in time. The actual visualization of actions of the human and the robot (verbal and non-verbal) has been done in Matlab. The figures were then compared and different timing and sequences of behaviors were identified. Hence, the visualization allowed different user strategies to be discovered. As visualization is a qualitative approach, the sequences that can be analyzed are restricted to a number that is adequate to still have the overview of the data.

Table 3-11. Overview of visualizations

Object-teaching 1	no visual analysis
Object-teaching 2	no visual analysis
Home tour 1	no visual analysis
Home tour 2	visual analysis of the users' strategies to attain the robot's attention (Section 5.2.2)

3.5 Off-talk analysis

The methods introduced so far focus on the interaction between the user and the robot and much information can be gained with their help. However, also what happens next to the HRI can lead to even more insights or underline what the HRI analyses suggest. One such source of additional information is off-talk. According to Batliner, Hacker, and Nöth (2006), users do behave naturally in interaction with more elaborated automatic dialog systems. That is why also phenomena such as speaking aside occur. This phenomenon is often called off-talk. Oppermann, Schiel, Steininger, and Beringer (2001) define off-talk as talk that is not directed to the system such as talk to another person, talk to oneself, or reading aloud. Off-talk is usually a problem for systems because they cannot tell it apart from system-directed talk. In most cases the system should not react to the utterances or process them in a special way, for example, on a meta-level as a remark about a problem in the interaction.

In the home tour studies many situations occurred in which the participants produced utterances that were not directed to the system. Most of them were questions or remarks to the experimenter. They were evident during the training phase and between the training phase and the actual trial. During the trial, communication with the experimenter was avoided whenever possible. That is why almost no off-talk utterances were produced in this phase of the study. However, the utterances from the other phases were found to be very useful, telling much about the users' conception of the situation. Therefore, they were included in the analysis below wherever it seemed adequate.

Many of the utterances were questions regarding the order of events in the study like "What do I have to show now?"; "Do we go to the dining room now?"; "Which one is the dining room?". These utterances were not taken into account in the analysis, as they do not concern the actual interaction with the robot and the expectations of the users. Nevertheless, they will be kept in mind for future study design.

The off-talk utterances were also annotated in ELAN and evaluated manually since the number of utterances was manageable without further analysis or categorization. Altogether, 20 off-talk utterances that were actually connected to the interaction with the robot were found. The utterances will be cited wherever they contribute to the evaluation of the situation and the users' expectations. They will not be summarized in a single section because, as stated above, they rather serve to underline the findings of other analyses.

Table 3-12. Overview of off-talk analyses

Object-teaching 1	no off-talk analysis
Object-teaching 2	no off-talk analysis
Home tour 1 and 2	conjunct off-talk analysis

3.6 Questionnaires and interviews

In contrast to off-talk that occurs during interaction, questionnaires and interviews were applied to get the participants to reflect on their experiences after the interaction. In the following, results of questionnaires and interviews from the home tour study will be analyzed. The focus is not on comparing the two trials but on presenting an overall impression and its implications for the users' expectations.

The main goal of the questionnaires was to get an overall evaluation of the robot and the interaction (see Appendix C). They included items on liking of the robot and judgments of its behavior (talkativeness, speed, interest, politeness, etc.). Moreover, they contained questions on how hard it was to handle the robot (ease of use, frustration, degree of concentration needed to interact with the robot, etc.). All of these items represent the participants' experience in numbers. To get a clearer idea about what was most impressive (positive and negative) about the interaction, the participants were interviewed. The interviews followed guiding questions about special impressions of the interaction, positive and negative situations in the interaction, and impression of speech recognition and output (see Appendix D). Moreover, the participants were asked what it was about the robot that caught their attention and where they looked to get information. The results of the questionnaires and interviews of the home tour are presented in Section 5.4.

Table 3-13. Overview of questionnaires and interviews

Object-teaching 1	questionnaire results in part published in Section 4.1, all results published in Lohse, Rohlfing, Wrede, and Sagerer (2008), no interviews
Object-teaching 2	no questionnaires, no interviews
Home tour 1 and 2	conjunct analysis of questionnaires and interviews (Section 5.4)

3.7 Overview of the methods for HRI data analysis and their purpose in the analysis process

Figure 3-9 gives an overview of the methods that were developed and are used in the following. It names the goals of the methods, describes which kind of data they are based on, and provides a short description.

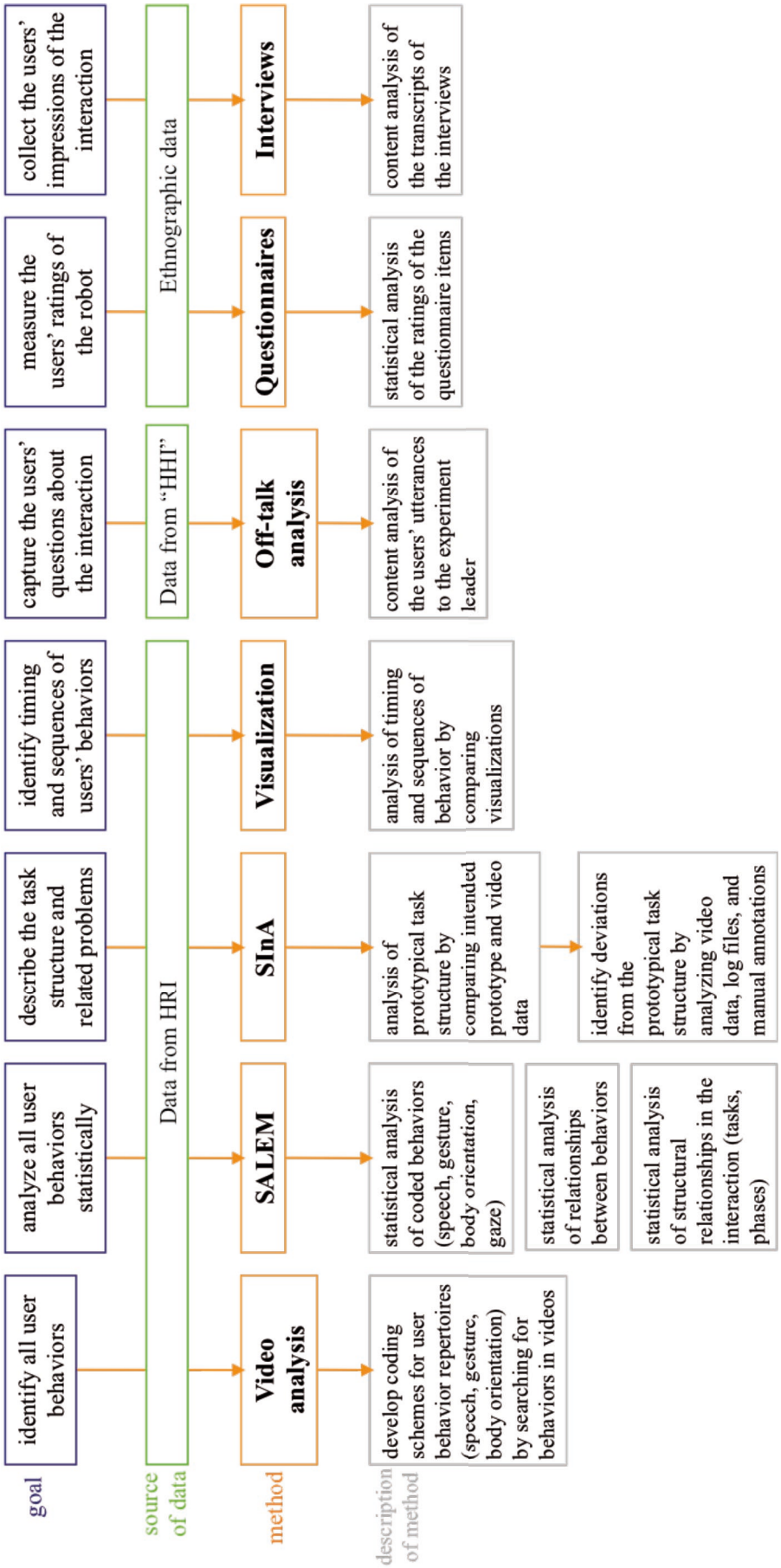


Figure 3-9. Overview of methods

4 HRI data analysis of the object-teaching studies

In the current chapter, the presentation of the results of the data analysis begins with the object-teaching studies because these are embedded in a more specific context than the home tour. They focus on one task, teaching objects to the robot, and are restricted to one spatial arrangement. In other words, both studies share the same physical social situation and are similar with respect to environmental setting, roles, rules, and tasks (see Section 2.1.1.3). Moreover, all participants interacted with the robot BIRON that looked similar despite having different abilities in the two studies. The goal of the chapter is to describe this object-teaching situation and to relate the results to the expectations of the users. The methods used in the context of the two studies are restricted to the analysis of user behavior with SALEM.

4.1 Object-teaching study 1

Objective

A first object-teaching study was designed to evaluate how and when users change their discursive behavior when interacting with a robot in an object-teaching task. The main question was what verbal and nonverbal behaviors and strategies the users apply. To research this question, data-driven coding schemes were developed for speech and gesture.

Subjects

The study was conducted with 15 participants who were German native speakers (13 female, two male) (see Section 1.4.1). All participants were students. Their age ranged between 22 and 37 years with an average of 24.7 years. Even though the majority had some experience working with computers (mean 3.3 on a scale of 1 [no experience] to 5 [very much experience]), only one person indicated that she had some minimal experience interacting with robots.

Procedure

The participants were provided with an introduction regarding their task and the robot BIRON (see Appendix E). Thereafter, they had to teach about five objects to the robot. The subjects could choose between different everyday objects, for example, a bottle, a cup, and a book.

During the study, the robot operated in its fully autonomous mode, which was necessary to produce realistic communication sequences including problems caused by the complex interaction of the diverse perceptual system components. Only speech recognition was simulated by manual text input to improve the recognition rate and to speed up the robot's reactions. The operator typed in all user utterances directed to the robot. Speech errors of all kinds (for example, cutting of words, vocalizations like "hmm") were ignored.[12] All utterances were typed in like they would occur in written language. The participants did not notice the operator during the interaction with the system. Next to the interaction, they were asked to fill in two questionnaires, one before and one after the interaction. The results have been reported in Lohse (2008).

[12]For a categorization of speech errors see Argyle, 1969, p.112.

Coding

In order to develop the coding schemes, all data from the trials were annotated with the help of ELAN. Annotations were made of

- verbal utterances of the human
- verbal utterances of the robot
- gestures of the human
- objects shown to the robot

The annotations were then analyzed and behaviors (speech and gesture) of the human were identified and grouped. For speech this was done with the help of a linguistic analysis. Accordingly, utterances with a similar grammatical structure and/or content were identified. Analyzing the videos, only units of speech and gesture that convey meaning concerning the task were taken into consideration. Thus, utterances like "mhm" or the scratching of the chin were not interpreted as conscious behaviors that were conducted to teach an object to the robot. With the help of the video analysis, eight task-related verbal behaviors were identified:

1. naming object (whole sentence) ("This is a cup.")
2. naming object (one word, very short utterance) ("Cup")
3. describing the object ("The cup is blue and has a handle.")
4. asking for feedback regarding the object ("BIRON, do you know what this is?")
5. asking for BIRON's general abilities and knowledge ("BIRON, what can you do at all?")
6. asking for BIRON's ability to listen/speak ("BIRON, can you hear me?")
7. asking for BIRON's ability to see ("Can you see the object?")
8. demanding attention for the user/object/task ("BIRON, look at me.")

While the first four behaviors describe object-related utterances, the last four include utterances about the abilities of the robot and its attentiveness. This shows that also in task-driven interaction, it seems to be important that the repertoire includes behaviors to find out what the interaction partner can do and whether it is attentive. In HHI these behaviors might be subtler than verbal utterances, because based on experience one can more easily estimate what abilities other humans might have and conclude from certain cues if they are attentive or not. The users often have less knowledge about the robot and the robot provides less feedback with this respect.

Next to typical verbal behaviors the data also implied some patterns concerning task-related gestures that the subjects used. Nine types of gestures were proposed:

1. Presenting the object
2. Moving the object once (up, down, to another position, rotate)
3. Moving the object continuously (back and forth, up and down, to different positions, rotate back and forth)
4. Moving the object closer to the robot

5. Manipulating the object (open the book/bottle)
6. Looking at the object
7. Pointing at the object
8. Imitating actions that can be performed with the object (drinking, eating, reading, etc.)
9. Holding the object

It could be questioned whether this categorization is exhaustive for all object-teaching tasks. This will be tested in part with a second corpus of data that is presented in Section 4.2.4. Moreover, when this coding scheme was developed, all behaviors of the users were included that seemed important at this point of time. That is why it contains holding the object (behavior 9). Technically, one could argue that this behavior is not a gesture in the sense it was defined in Section 3.2.2 because it has no sharp onset and offset and even more importantly it is not directed at the robot. This was taken into account in the analysis of the second object-teaching study as will be described below. There, the categorization presented here will be used as the basis for an adapted coding scheme.

Results

Even though the coding scheme shall here be presented as one main outcome of the first study, some results that were identified with the help of the coding schemes shall be briefly summarized. These results concern the change of the strategies by the users. Most changes were connected to situations when BIRON said that it had not understood or it could not do something. When this happened, the subjects tried to paraphrase, i.e., they switched between saying a whole sentence (behavior 1) and saying one word or a very short phrase (behavior 2). Another important reason for changing behavior was the need to verify if BIRON had understood something. This happened when the robot signaled that it understood and the users wanted to be sure if this was true. Hence, the participants asked for feedback (behavior 4), knowledge, and abilities (behavior 5) of the system. Another situation that caused the users to switch between behaviors was a missing reaction by the robot. When BIRON had not done anything for some time, the subjects started naming the object in a detailed manner (behavior 1) or describing the object (behavior 3). Last but not least, the participants changed their behavior when they showed a new object to the robot. In this case, they usually asked BIRON for attention (behavior 8) and named the object in a whole sentence (behavior 1).

When the users started another behavior was also analyzed with respect to gestures. Five typical situations during which the users switched between different gestures were identified. Primarily, the participants applied another behavior when a new object was chosen. Usually the object was then presented to the robot (behavior 1), the object was moved in front of the robot (behavior 2, 3), or the subjects pointed at the object (behavior 7). All these behaviors seemed to be applied to attain the robot's attention. Thus, in this situation the gestures seemed to have the same function as the speech, where asking BIRON for attention (behavior 8) was found to be most common. Similar behaviors were evident when the users tried to present the same object one more time because BIRON had not recognized it or had not done anything for quite some time. As described above, when BIRON had not understood something the users paraphrased. While

doing this, they also tried two different types of gestures. They held the objects (behavior 9), which often seemed to be a sign of disappointment. Some chose the opposite behavior though and tried to regain BIRON's attention by moving the object to another position (behavior 2). This might be due to the fact that the users felt that BIRON might not have seen the object at the previous location. The same new behaviors were chosen when BIRON had not done anything for quite some time. The last situation that typically caused a change in behavior was the description of an action (for example, "This is a pencil. It is used for writing."). In this case, a very close coherence of speech and gestures could be seen because the actions were described verbally and at the same time imitated in the gestures. The most common switches of gestures took place between presenting the object and moving it to another position. Thus, there was a constant change between holding the object still for the robot to recognize and trying to obtain the robot's attention.

All these switches in behavior showed that the participants conducted them in reaction to the robot's behavior. However, changes in user behavior seemed to be carried out consciously only when robot feedback for a certain channel was available. Thus, mainly changes in speech were reported by the participants when they were asked after the interaction how they adapted their behavior to the robot (see Table 4-1). All but two subjects only mentioned conscious linguistic adaptations. The participants did not consciously adapt their gestures to the robot, even though a notable number of gesture behaviors were found and changes between them occurred as frequently as linguistic changes. These changes of the users' gesture behavior seemed to be an unconscious variation instead of an adaptation because there was no way for the users to find out which behavior was beneficial for the robot.

Table 4-1. Adaptation behaviors reported by the participants in the first object-teaching study (# = number of participants that mentioned the adaptation behavior)

Adaptation behavior	**#**
Manner of speaking	8
• speak more clearly	2
• vary intonation	1
• vary loudness	2
• speak more slowly	3
Sentence structure	17
• verification questions	4
• switch between different sentence structures	1
• simple sentences	2
• imperative sentences	2
• one-word sentences	6
• special sentence structure ("This is a…")	2
Content of utterances	7
• repetitions	3
• paraphrasing	1
• descriptions	3
Change between gestures	2
• hold object into the camera focus	1
• change between moving object and holding it still	1

= number of participants that mentioned the adaptation behavior

The most common conscious adaptation was to use one-word sentences (sentences that only contain the name of the object taught to the robot). This finding implies that the users thought that BIRON only understood very simple sentences. Another common behavior was to ask verification questions. These questions show the users' need for more feedback.
Moreover, it was found that feedback influenced the expectations of the users during the interaction insofar as their views of the speech in- and output of the robot were rather consistent after the interaction, whereas they were not sure about BIRON's abilities to recognize people, mimic, and gestures. As the feedback seems to be such an important factor, in the following it shall be evaluated whether the coding scheme holds for interaction with more sophisticated robot behaviors and if there are differences when the robot behavior is varied systematically.

Conclusion

The analysis of the first object-teaching study leads to some implications with respect to expectation theory and the model presented above. The consistent view on speech in- and output across the participants after the interaction points to the fact that they developed target-based expectations during the interaction with the robot. Since the judgment was similar across subjects, it can be assumed that it was mainly influenced by the situation, in particular by the skills of the robot, and not by the personality of the users. In contrast, the participants did not agree on BIRON's abilities concerning recognizing people, mimic, and gestures. They perceived the situation differently with these respects because the robot did not provide them with explicit feedback. Hence, they did not form similar expectations. Also changes of user behaviors as a result of the robot's behavior were only conscious for speech but not for gestures. Speech behaviors changed when the robots' behavior disconfirmed the users' expectations. Thus, the robots' behavior directly influenced the users' behavior. Since the robot did not gesture, this relation was not established for the users' gesture behaviors which users changed unconsciously. Accordingly, the users' expectations strongly depend on the behavior of the robot with respect to the modalities.

4.2 Object-teaching study 2

The second object-teaching study was conducted as a follow-up to the first study. The scenario and the task stayed the same, which enables a comparison between both object-teaching studies and the usage of the coding schemes in this second study.

Objective

Next to gesture and speech, the data of the second study were also analyzed with respect to gaze behavior of the user, phases of the interaction, and differences between a positive and a negative trial. Moreover, while the analysis in the first study was mainly unimodal, this second study served to reveal multimodal relationships. Furthermore, it allowed for detailed descriptive statistics and transition probabilities of the results because the SALEM toolbox to calculate these statistics much faster and more accurately on ELAN files was just developed before the analysis of the second object-teaching study. Therefore, this study does not only display the differences in study design but also the advancements of the analysis approach.

Subjects

The study was conducted with eleven German native speakers (five female and six male) ranging from 22 to 77 years in age, nine of whom had never interacted with a robot before. This sample can be seen as an improvement compared to the first study because the age range was much bigger and not all the participants were students.

Procedure

The participants received a similar introduction to the robot BIRON and the task as in the first study (see Appendix F). In contrast to the first study, the robot was now controlled by a wizard (which the subjects did not know) and the participants had to complete two trials: a *positive* one where BIRON termed most of the objects correctly, and a *negative* one where BIRON misclassified the majority of the objects. Section 4.2.1 introduces how both trials differed and proves that the robot (i.e., the operator) actually performed better in the positive trials. One session lasted about ten minutes. Between the sessions, the objects were exchanged to make the subjects believe that the recognition performance of the robot was to be evaluated on another object set.

4.2.1 Differentiating positive and negative trials

One main aspect in this second object-teaching study was to find out whether participants behave differently depending on the course of the interaction. Thus, all users had to complete two trials. In one of them the robot recognized almost all objects and repeated their names when asked for them. In the other trial the robot failed to recognize the objects. The trials were counterbalanced to avoid sequence effects. In fact, the analysis did not show such effects and it did not make a difference whether a person completed the positive or the negative trial first.

To differentiate the trials and to determine whether they were in fact different from each other, all sequences of teaching objects to the robot were coded as belonging to one of the categories success, vague, clarification, failure, problem and abort. Some sequences could not be categorized clearly because they were a success because the robot did what it was supposed to do but the person did not perceive it as such or, vice versa, the robots' reaction was thought to be vague but the participants interpreted it as a success. These cases were not taken into consideration here. The outcomes of the sequences were distinguished as follows (utterances were translated from German to English by the author):

- success: BIRON names the object correctly
- failure: BIRON names the object incorrectly
- problem: BIRON does not say an object name at all but utters:
    - "I don't know the object"
    - "I don't know the word"
    - "I don't know"
- vague: BIRON's utterance is vague and does not include the object name
    - "I have seen the object"
    - "This is interesting"
    - "I really like it"

- clarification: BIRON asks the user to clarify:
  - "Pardon?"
  - "I didn't completely understand you. Could you please repeat that?"
  - "Have you shown me the object before?"
- abort: BIRON does not answer at all and the user interrupts the sequence

The attempts to teach an object were coded accordingly leading to the results presented in Table 4-2. Far more success cases were coded in the positive trials than in the negative ones (45.39% of all attempts compared to 19.78%). The percentage of successful teaching sequences that each person achieved in the positive trials was significantly higher than the percentage of failures (T-test (two-tailed), df=10, T=7.766, p=.000**). Also the percentage of failures differed significantly, being higher in the negative trials (15.13% compared to 39.55%; T-Test (two-tailed), df=10, T=7.671, p=.000**). Moreover, the users only aborted the interaction in the negative trials which is a conclusive result because this was the last option for the participants if the interaction got stuck. The percentage of clarifications, problems, and vague cases did not differ between the trials. This was positive because these cases also occur in many situations in HHI; in other words, for example, clarification questions occur regularly and are no sign for a negative interaction.
Altogether, it can be concluded that both conditions could clearly be distinguished from each other and are a basis for the comparison of two HRI situations that differ regarding the success of the object-teaching task. With respect to the model, the two conditions should lead to different behaviors of the users caused by the influence of the robots' behaviors on their expectations.

Table 4-2. Outcomes of the object-teaching sequences in the positive trials, negative trials, both trials

outcome	**positive trials**		**negative trials**		**both trials**	
	count	%	count	%	count	%
success	123	**45.39**	88	19.78	211	29.47
vague	8	2.95	12	2.70	20	2.79
clarification	66	24.35	122	27.42	188	26.26
failure	41	15.13	176	**39.55**	217	30.31
problem	33	12.18	38	8.54	71	9.92
abort	0	0	9	2.02	9	1.26
sum	271	100	445	100	716	100

4.2.2 Differentiating phases of the interaction

For further evaluation, the object-teaching attempts were subdivided into four phases (Lang et al., 2009):

1. present: the subject presents the object to BIRON and tells its name or asks for the name
2. wait: the subject waits for the answer of the robot (not mandatory)
3. answer: the robot answers
4. react: the subject reacts to the answer of the robot

Table 4-3. Descriptive statistics of the phases of the interaction in the object-teaching task

phase	count	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)	overall duration (seconds)	duration of all phases (%)
present	734	.32	8.96	2.03	1.10	1.82	1490.02	15.10
wait	724	.06	51.55	**6.75**	5.59	5.43	4887.00	**49.52**
answer	725	.36	9.19	1.77	1.04	1.56	1283.25	13.00
react	729	.39	20.91	3.03	2.45	2.33	2208.87	22.38
all	2912	.06	51.55	3.39	3.72	2.05	9869.14	100

Figure 4-1. Sequence of phases in the object-teaching task

These phases represent different situations of the interaction that might be connected to certain behaviors. For example, if the participants use certain phrases only in the react phase, this can be a sign for the robot of how well the interaction is going, how the situation is perceived by the user, and which expectations the user has in the situation. Therefore, the phases are used to analyze differences in the users' behavior repertoires in certain interaction situations. To evaluate the structure of the phases, they were analyzed with SALEM regarding their descriptive statistics which are presented in Table 4-3 and depicted in Figure 4-1.

The analysis reveals that the number of occurrences (count) is similar for all phases. One needs to keep in mind though, that the react phase can also be a presentation which led to an overlap of 36.55% between the two phases. All other phases do not or only barely overlap. Moreover, the table reveals that the participants spent most time waiting because the mean duration of these instances is by far the longest (even though the number of occurrences is similar to all other phases). The mean durations of presentations and answers of the robot were shorter. This might be due to the actual functional differences between the phases. Further analysis will reveal whether this actually shows in the behavior of the users (see Section 4.2.3). For now, it shall be noted that no differences in the phase structure were found between the positive and the negative trials, no matter in which order they were presented. The users spent about the same percentage of time per trial in each phase and the mean length of the phases was also similar. This was equally true for the wait phase. Consequently, if there are any differences between the positive and the negative trials, they are due to what the robot said and not to how long it needed to reply.

4.2.3 Analysis of speech in the object-teaching study

The first modality that will be analyzed in this section is speech. As has been explained before, speech here does not refer to speech recognition, prosody, dialog management, or all sorts of errors that can take place in spoken HRI. All these are special fields with a wide range of literature. Rather, the term speech here addresses utterances of the human. The main questions are what was said and what was likely to be said next in order to conclude why it was said. This approach is human-centered. That is why the speech of the robot is of minor importance here.

Coding

To analyze the participants' speech repertoire in the object-teaching task, part of the data were annotated using the coding scheme developed in the first object-teaching study which shall be replicated here:

1. naming object (whole sentence)
2. naming object (one word, very short utterance)
3. describing the object
4. asking for feedback regarding the object
5. asking for BIRON's general abilities and knowledge
6. asking for BIRON's ability to listen/speak
7. asking for BIRON's ability to see
8. demanding attention for the user/object/task

However, it quickly turned out that quite a few changes had to be made to the scheme because it did not cover all utterances in the new study, and new research findings and questions revealed short-comings. Therefore, i.e., describing the object (behavior 3) was divided into two behaviors: describing the object including object name and describing the object without using the object name. This was done because when naming the object the utterance alone is sufficient to teach the object name; this is not true when the name is not part of the utterance (see Iverson, Longobardi, & Caselli, 1999; Rohlfing, to appear). Behaviors that do not name the object must, therefore, be used as an add-on to other behaviors that include the object name. Of course, also phrases that include the object name can succeed each other. Another change was to add a code for utterances like "this" ("das da") which mainly accompany gestures in a reinforcing manner, because without the gesture the meaning of the verbal utterance would not become clear.
In the process of coding, also two more behaviors were identified in the category of utterances toward the object: comment an action that is performed with the object such as "I will put the object here" and comments on the object name such as "This is a long word". No new behaviors were identified with respect to utterances towards BIRON and the question for attentiveness. They were transferred to the new coding scheme without changes. However, a whole new category needed to be added that did not play a role in the first trial. It includes utterances about the interaction (behaviors 21-26). All these utterances have a positive and a negative counterpart. They are reactions to the robot's recognition of an object, answers to the robot's questions, for example, to repeat an utterance and further comments about the robot's utterances which are not questions and do not display recognition of an object. Finally, a category was added to code all utterances that did not fit in the other categories. This category was used as few times as possible. The following coding scheme resulted from all these changes:

Utterances about the object:

1. naming object (whole sentence), ("This is a cup")
2. naming object (one word, very short utterance), ("Cup")

3. describing the object (shape and/or function) and naming it (“The cup is blue and has a handle”, “A book to read”)
4. describing the object (shape and/or function) without naming it (“to read”)
5. asking for feedback regarding the object (“BIRON, do you know what this is?”)
6. deictic words (“this”)
7. comment on action that is performed with the object (“I put this in here”)
8. comment on object name (“This is a long word”, “a new word”)

Utterances about BIRON:

11. asking for BIRON’s general abilities and knowledge (“BIRON, what can you do at all?”)
12. asking for BIRON’s ability to listen/speak (“BIRON, can you hear me?”)
13. asking for BIRON’s ability to see (“Can you see the object?”)

Demand for attention:

14. demanding attention for the user/object/task (“BIRON, look at me”, “BIRON”, “Hello”)

Utterances about the interaction:

21. praise because BIRON has correctly recognized the object (“Right”; “Exactly” “Good”)
22. commenting wrong robot utterances (“No”; “This is not a cup”)
23. positive answer to robot’s question (“Yes” after being asked to repeat an utterance)
24. negative answer to robot’s question: (“No” after being asked to repeat an utterance)
25. positive comment towards robot/robot utterance which is not a question/a sentence in connection to object recognition (“You are also interesting” as reaction to the robot utterance “This is interesting.”)
26. negative comment towards robot/utterance which is not a question/a sentence in connection to object recognition (“No, left from my point of view” after the robot has turned to the wrong direction)

31. other (all utterances that do not fit in the above categories)

Results

1633 user utterances were coded with this scheme. Only very few (8, 0.49%) belonged to the category “other” (31). Altogether, the participants spoke 19.5% of the time in the four phases.[13] Since the analysis was restricted to the phases, utterances in the context of greeting and farewell were not coded, even though these activities were part of the procedure.

[13]For comparison: 937 robot utterances were coded in the four phases (735 in the answer phase, 177 in the react phase, 18 in the present phase and seven in the waiting phase). The robot spoke 15% of the time because the mean duration of its utterances was longer (1.6 seconds). However, this finding could be influenced by the fact that the robot utterances were not coded and segmented manually but consisted of sentences as specified in the script.

Table 4-4. Descriptive statistics of the speech behaviors

code	count	count of all annotations (%)	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)
1	530	**32.46**	0.39	8.72	1.88	0.82	1.73
2	248	**15.19**	0.40	4.33	1.31	0.52	1.34
3	33	2.02	1.22	6.42	**2.64**	1.12	2.47
4	39	2.39	0.70	6.73	1.94	1.28	1.57
5	155	**9.49**	0.45	2.59	1.16	0.39	1.04
6	23	1.41	0.34	2.19	0.90	0.53	0.67
7	4	0.24	1.07	1.42	1.26	0.16	1.27
8	4	0.24	0.91	2.60	1.52	0.77	1.29
11	10	0.61	1.08	3.26	1.76	0.72	1.55
12	1	0.06	1.27	1.27	1.27	0	1.27
13	29	1.78	0.69	3.42	1.44	0.70	1.11
14	32	1.96	0.51	3.19	1.16	0.65	0.92
21	257	**15.74**	0.24	3.67	0.93	0.48	0.82
22	184	**11.27**	0.16	4.21	0.88	0.51	0.77
23	41	2.51	0.20	3.22	0.84	0.58	0.75
24	15	0.92	0.40	2.53	0.86	0.55	0.73
25	9	0.55	0.66	1.55	1.05	0.30	1.01
26	11	0.67	0.56	2.11	1.30	0.51	1.17
31	8	0.49	0.97	3.14	1.84	0.91	1.37
all	1633	100	0.16	8.72	1.40	0.80	1.24

Having coded the utterances with the help of the coding scheme enabled the analysis with SALEM. The questions that drove the analysis were: How many times were certain types of utterances used?; What were the transition probabilities between the behaviors?; What utterances were used in the phases of the interaction?; Were there differences between the positive and the negative trials? I.e., does the usage of the utterances change when the interaction runs poorly? Table 4-4 presents the results for all utterances regardless of phases and conditions.

Naming the object in a whole sentence (behavior 1) was the most common behavior for the task. In 32.46% of the cases, the participants labeled the object in a whole sentence. Also short utterances (behavior 2) were commonly used to introduce the objects (15.19% of all utterances). Moreover, praise and comments on wrong robot utterances (behaviors 21 and 22) occurred frequently (15.74% and 11.27% of the utterances, respectively). This shows that the users very commonly praised the robot but also told it if it had done something wrong. Further analysis will show how these behaviors were distributed between the positive and the negative trials. The users also frequently asked the robot for feedback regarding the object (9.49%) to make sure whether it had really understood the object name and remembered the object later in the interaction. This behavior was certainly connected to the instruction to verify if the robot had actually learned the object.

On the other end of the scale, some behaviors were only annotated few times, especially asking the robot for its ability to listen/speak (behavior 12) was only used once. This behavior was

much more prominent in the first study which was probably due to the fact that the robot in this study often needed a long time before it replied something to the users' utterances. Therefore, the users were very insecure about its abilities. In contrast, in the second study the robot was controlled by an operator who reacted rather quickly. Consequently, this case did not occur. However, the behavior might still be important in autonomous interaction.

Also comments on actions that are performed with the object (behavior 7) and comments on object names (behavior 8) were rarely used. However, the codes should not be changed until they have been evaluated with data from other studies where they might be more prominent.

In the following, the behaviors were analyzed in groups (utterances about the object [behaviors 1-8], utterances about BIRON [behaviors 11-13], demand for attention [behavior 14], utterances about the interaction [behaviors 21-26], and other [behavior 31]) (see Table 4-5).

The table illustrates that utterances about the objects were most common (63.44% of all utterances) which is in line with the task. The users made clear what task they wanted to accomplish. Utterances about the interaction were carried out half as often (31.66% of all utterances). Utterances about BIRON were not common (2.45% of all utterances). This shows that the users focused on the task and not on getting to know the robot itself. Instead of asking how it can learn about objects they simply introduced the objects. This finding is certainly influenced by the task but it could also change with the performance of the robot. Therefore, now the positive and the negative trials are compared. The results for the groups are depicted in Table 4-6 and Table 4-7. Only the results for the groups are shown because the single behaviors hardly differed between the conditions. However, some interesting discoveries were made. Most striking were the differences between the behaviors 21 and 22 (praise because BIRON has correctly recognized the object, commenting wrong robot utterances). Praise occurred significantly more often in the positive trials (23.64% vs. 10.82%; T-Test (two-tailed), df=10, T=11.647, p=.000**) while significantly more comments about wrong utterances were counted in the negative trials (14.80% vs. 5.59%; T-Test (two-tailed), df=10, T=8.690, p=.000**). This finding is in accordance with common sense. Anyhow, it signals that if the robot could differentiate between praise and negative comments, it could better understand whether the interaction was going well or not.

The differences with respect to questions for attention (behavior 14) were not quite as clear (1.44% in the positive trials vs. 2.28% in the negative trials), even though one could expect the users to ask for attention more often if the interaction is not going well. However, the behavior was not as prominent in both trials and some participants did not use it at all. This might be due to the fact that the robot was teleoperated and reacted within an appropriate time in all cases. Therefore, it was not possible to identify significant differences in the mean percentage of usage of this behavior. The same was true for questions about BIRON's ability to see (behavior 13), but there was also a trend that it was used more often in the negative trials (0.80% in the positive trials vs. 2.38% in the negative trials). Thus, it seems that the participants ask more about the robot itself and more often demand its attention when the interaction is problematic.

Another behavior that they tended to show more often in the negative trials was naming the object in one word or a very short utterance (behavior 2) (12.14% vs. 17.08%). This trend was not consistent for all participants though and the mean percentage of usage did not differ

significantly. Anyhow, seven of the eleven participants produced a higher percentage of this behavior in the negative trials.

One of the questions posed here is whether there were typical sequences of behaviors. To answer this question, transition probabilities[14] were calculated. Altogether, it was found that for the number of cases analyzed here (1633), the transition probabilities rarely exceeded 20%. This finding is in line with the literature on HHI that states that transition probabilities are often low because usually interaction is too complex to be predictable by changes from one state to

Table 4-5. Descriptive statistics of groups of speech behaviors

utterance about:	count	count of all annotations (%)	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)
object	1036	**63.44**	0.34	8.72	1.63	0.82	1.49
robot	40	2.45	0.69	3.42	1.51	0.70	1.26
attention	32	1.96	0.51	3.19	1.16	0.65	0.92
interaction	517	**31.66**	0.16	4.21	0.91	0.50	0.81
other	8	0.49	0.97	3.14	1.84	0.91	1.37
all	1633	100	0.16	8.72	1.40	0.80	1.24

Table 4-6. Descriptive statistics of the speech behaviors in the positive trials

utterance about	count	count of all annotations (%)	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)
object	390	62.30	0.34	8.72	1.70	0.93	1.54
robot	9	1.44	0.69	2.36	1.19	0.47	1.11
attention	9	1.44	0.51	1.41	0.97	0.26	0.93
interaction	216	34.50	0.22	3.22	0.90	0.47	0.81
positive	170	**27.16**	0.24	3.22	0.92	0.49	0.82
negative	46	**7.35**	0.22	2.11	0.82	0.39	0.74
other	2	0.32	2.53	3.04	2.79	0.36	2.79
all	626	100	0.22	8.72	1.41	0.88	1.22

Table 4-7. Descriptive statistics of the speech behaviors in the negative trials

utterance about	count	count of all annotations (%)	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)
object	646	64.15	0.40	7.92	1.59	0.75	1.47
robot	31	3.08	0.73	3.42	1.61	0.74	1.40
attention	23	2.28	0.52	3.19	1.24	0.74	0.91
interaction	301	29.89	0.16	4.21	0.92	0.52	0.82
positive	137	**13.60**	0.20	3.67	0.92	0.50	0.82
negative	164	**16.29**	0.16	4.21	0.92	0.55	0.82
other	6	0.60	0.97	3.14	1.53	0.81	1.23
all	1007	100	0.16	7.92	1.38	0.75	1.24

[14] Probability that a particular kind of act will be followed by another (Argyle, 1988)

another (Argyle, 1988). For example, at times, earlier actions influence later actions, a case that cannot be depicted by the transition matrix. Moreover, structural analysis does not include any information about causes and effects of social events. However, the data offer the possibility to analyze transition matrixes for special situations which increases the transition probabilities considerably. Therefore, here the transition probabilities are presented for the cases that were annotated as success, failure or clarification (see Section 4.2.1). These cases are the ones that took place most often and summed up to 86% of the attempts to teach an object to the robot. The tables (Table 4-8, Table 4-9, and Table 4-10) only depict the speech behaviors that were carried out most often in order to increase the clarity of the results.

Table 4-8. Successor transition matrix for the teaching sequences with the outcome "success"

		predecessor			
		1	2	21	5
successor	1 (naming object, whole sentence)	0.0909	0.0597	**0.4350**	0.0727
	2 (naming object, short utterance)	0.0839	0.1194	0.1794	0.0182
	21 (praise)	**0.5664**	**0.7612**	0.1031	**0.6727**
	5 (asking for feedback regarding the object)	0.0699	0.0149	0.1211	0.2182
	count	143	67	225	55

Table 4-9. Successor transition matrix for the teaching sequences with the outcome "failure"

		predecessor			
		1	2	22	5
successor	1 (naming object, whole sentence)	**0.3118**	0.2448	**0.5922**	0.1389
	2 (naming object, short utterance)	0.1749	0.2378	0.2185	0.0556
	22 (comment on wrong robot utterance)	**0.3497**	**0.3636**	0.1019	**0.5833**
	5 (asking for feedback regarding the object)	0.0380	0.0490	0.0437	0.1388
	count	264	143	207	36

Table 4-10. Successor transition matrix for the teaching sequences with the outcome "clarification"

		predecessor				
		1	2	22	23	5
successor	1 (naming object, whole sentence)	**0.3962**	0.3281	**0.5806**	**0.6591**	0.1905
	2 (naming object, short utterance)	0.2358	**0.4123**	0.3548	0.1363	0.0476
	22 (comment on wrong robot utterance)	0.0613	0.0763	0	0.0682	0.0476
	23 (positive answer to robot's question)	0.0896	0.0611	0	0.0909	0.1190
	5 (asking for feedback regarding the object)	0.0566	0.0305	0.0645	0.0227	**0.3095**
	count	212	133	31	44	42

These behaviors sum up to 86.12% (success), 88.80% (failure), and 86.52% (clarification) of all utterances (569, 732, and 534 utterances, respectively). They also cover more than 80% of the transitions for all cases but one (behavior 5 in the matrix for the outcome "clarification"). Often they are higher than 90%. In the following, the insights that they provide are compared with the results from the first object-teaching study that were presented in Section 4.1.

In the first object-teaching study, the participants most commonly switched from naming the object in a whole sentence (behavior 1) to naming the object in a very short utterance (behavior 2) if the robot said that it had not understood or that it could not do something that the user had asked for. Switching between these behaviors was also found very often in the second study (19.7% for all utterances). However, even more often the users repeated behavior 1 (24.0% for all utterances in all situations). They especially did this in the confirmation sequences (39.62%) and in the failure situations (31.18%). This is not contradictory to the first study because there the analysis only focused on changes and did not take cases into consideration when behaviors were repeated.

The second finding of the first study was that the users commonly switched from presenting the object in a whole sentence (behavior 1) to asking for feedback about the object (behavior 5) and asking for knowledge and abilities (behavior 11). Here, the question for feedback was only found to be preceded by a presentation (behaviors 1 and 2) in 22.6% of all cases. In the successful interaction, eleven presentation utterances were followed by a question for feedback which equals 20.0%. In the confirmation situation, 16 utterances of behaviors 1 and 2 resulted in the question for feedback (38.1%). In the failed interaction, 17 presentations were succeeded by a question for feedback (47.22%). This difference seems to be due to the fact that the users praised more than they said negative comments, because in the success cases typically the presentation was followed by praise which was succeeded by the question for feedback about the object, which then again was often followed by more praise. This sequence hardly ever occurred in the first user study because BIRON's feedback utterances were very limited. In contrast, the interaction resulted in the typical sequence of failed attempts: the users presented the object which the robot failed to recognize and the users presented the object again (a case that was not represented in the first study) or asked the robot about the object. Negative comments as a counterpart to praise also took place in this case but were less common. This might be a reason why the comments about the interaction were not as present in the first study. Another reason that this connection was not recognized in the first study could be that the comments about the interaction were just not important when analyzing that data and were not coded. One would have to analyze the data again with the new coding scheme and the new research question in mind to determine whether actually the content of the interaction was different or the coding of the data because of the research interest at that point of time. However, this is out of scope here.

Another comment on the interaction that was not reported in the first study was complying with BIRON's requests (behavior 23). This relation can be seen here in the confirmation case. If the robot asked for repetition, the users often first confirmed that they would repeat the utterance before they actually did so. Accordingly, a presentation (behavior 1) was followed by utterances

that signaled compliance with a request in 8.96% of the cases. These utterances were succeeded by a presentation with a probability of 65.91%.

In contrast to the first study, also the strong relation between presenting the object in a whole sentence (behavior 1) and asking for BIRON's abilities and knowledge (behavior 11) could not be confirmed. Altogether, behavior 11 only occurred ten times in the study presented here. This finding indicates that asking for abilities and knowledge is a behavior that the participants use much more frequently when the robot has fewer abilities and they try to find out what it can do at all. Asking for abilities thus points to the users' belief that the robot can do more as it has done so far.

In the first study, it was also found to be typical that the users switched from presenting the object in a whole sentence (behavior 1) to describing the object. Describing the object in the new coding scheme equals behaviors 3 (including object name) and 4 (without object name). Both behaviors were carried out only a few times in the second study (33 [2.02%] and 39 [2.39%], respectively). However, they were in fact often preceded by a presentation of the object (42.4% and 59.0%).

Finally, in the first study it was evident that the users changed their behavior when they presented a new object. This case will be discussed in the context of the analysis of speech in the phases. For now, it shall be concluded that the transition probabilities between utterances for a restricted task like teaching objects to a robot with limited feedback abilities, allow to discriminate success, failure, and confirmation situations in the task (at least on a theoretical basis) and, thus, to anticipate the users' behavior with a high probability.

While the results presented so far depended on the content of the interaction, the following analysis will show that also the structure of the interaction which is represented with the phases present, wait, answer, and react, influences the verbal behavior (see Table 4-11). To begin with, it has to be noted again that the react phase can overlap with the present phase if the reaction is a presentation. This is proven by the finding that 23.68% of the utterances in the react phase are presentations of the object in a whole sentence and 16.63% are presentations in a short utterance. This results in the fact that 51.25% of the presentations in long utterances (behavior 1) and 96.93% of the presentations in short utterances are reactions to a previous attempt to teach the object to BIRON. Additionally, this finding underlines that short utterances to present the object (behavior 2) occurred when the utterance was repeated and the users concentrated on the object name only.

Moreover, some utterances overlapped with two phases, i.e., they began in one phase and ended in another. This was rarely the case in the present phase and the react phase (2.48% and 1.50%, respectively). However, in the wait phase users had not finished talking when the robot answered in 11.5% or, vice versa, the users still talked in 72.4% of the cases (42 cases) when the answer phase began. This means that the robot often interrupted the users. On the other hand, the users also interrupted the robot because 24.1% of the utterances (14 cases) in the answer phase were made while the robot was still talking.

In the first object-teaching study, it was found that new objects were often presented in a whole sentence. This is also true for the second study. Assuming that all presentations in whole sentences that were not reactions were meant to introduce new objects, a number of 214 presen-

Table 4-11. Descriptive statistics of speech behaviors in phases

	present phase			wait phase			answer phase			react phase		
code	**count**	**%**	**mean dur. (sec.)**	**count**	**%**	**mean dur. (sec.)**	**count**	**%**	**mean dur. (sec.)**	**count**	**%**	**mean dur. (sec.)**
1	439	**49.55**	1.96	63	**22.58**	1.57	20	**34.48**	1.22	225	**23.68**	1.74
2	163	**18.40**	1.34	73	**26.16**	1.29	6	10.34	1.58	158	**16.63**	1.33
3	21	2.37	2.88	8	2.87	2.45	3	5.17	2.38	16	1.68	2.70
4	6	0.68	2.67	27	9.68	1.78	5	8.62	1.29	9	0.95	1.85
5	96	10.84	1.22	51	**18.28**	1.06	12	**20.69**	0.91	27	2.84	1.35
6	9	1.02	1.14	14	5.02	0.75	1	1.72	1.62	4	0.42	1.44
7	0	0	0	0	0	0	0	0	0	4	0.42	1.26
8	0	0	0	3	1.08	1.52	0	0	0	1	0.11	1.53
11	3	0.34	2.27	6	2.15	1.41	2	3.45	1.55	4	0.42	2.29
12	0	0	0	1	0.36	1.27	0	0	0	0	0	0
13	16	1.81	1.64	11	3.94	1.06	1	1.72	1.09	15	1.58	1.73
14	14	1.58	1.31	11	3.94	1.10	1	1.72	0.72	8	0.84	1.32
21	5	0.56	1.21	0	0	0	1	1.72	1.81	236	**24.84**	0.93
22	76	8.58	0.79	3	1.08	1.40	1	1.72	0.89	178	**18.74**	0.87
23	22	2.48	0.76	0	0	0	5	8.62	0.98	37	3.89	0.81
24	12	1.35	0.78	0	0	0	0	0	0	15	1.58	0.86
25	1	0.11	1.45	0	0	0	0	0	0	6	0.63	1.09
26	2	0.23	1.42	4	1.43	1.24	0	0	0	6	0.63	1.54
31	1	0.11	3.04	4	1.43	1.20	0	0	0	1	0.11	3.14
overall	886	100	1.62	279	100	1.35	58	100	1.25	950	100	1.25

tations were made in whole sentences (behavior 1). This number is much higher than the number of short presentations (5) (behavior 2).

The main question here was to determine whether the users' utterances differed in the phases. In this context, a basic measure is to differentiate how much the users talked. This measure depends on the coding of the phases. The coding convention clearly defined that the present phase was almost restricted to the users' utterances when presenting an object. Thus, the participants talked 96.27% of the time. In the wait phase and the answer phase they only spoke 7.72% and 5.65% of the time, respectively. In the react phase, the utterances covered 53.89% of the time. This number was lower than in the present phase because the reactions were not restricted to speech but also included facial expressions and other nonverbal behaviors. In consequence, the amount of human speech differed between the phases. Now the question is whether also the content differed. As for the comparison of speech in sequences with different outcomes, this comparison shall be restricted to the most common behaviors. These again are presenting the object in a whole sentence (behavior 1), presenting the object in a short utterance (behavior 2), asking for feedback regarding the object (behavior 5), praising the robot because it has correctly recognized the object (behavior 21), and commenting a wrong robot utterance (behavior 22).

With respect to behavior 1, Table 4-11 shows that this behavior was most common in the present phase where it was counted in almost half the utterances. It was also used in the other phases but less frequently. The results differ for behavior 2. Short utterances introducing an object are more commonly used in the wait phase than in the present phase and less frequently in the other phases. Behavior 5 again has another distribution. It was most common in the answer phase and in the wait phase. It was less common in the present phase even though questions like "BIRON, what is this?" were also coded as presentations. Although these questions occurred most often in the present phase they were less important because other behaviors were used more often. Asking for feedback about the object was not common at all in the react phase (only 2.84%). What the users did in this phase was praise the robot (behavior 21) and comment on wrong robot utterances (behavior 22). In fact these two behaviors characterize this phase. They did not take place nearly as often in the other phases. Some comments on wrong robot utterance were coded in the present phase. This is in line with the coding scheme, where utterances like "No. this is a cup" were segmented in a comment on a wrong robot utterance (behavior 22) and a presentation of the object in a whole sentence (behavior 1). The react phase included both these utterances.

Conclusion

As a result from this analysis, the four phases can be differentiated and described as follows:

- present phase: many user utterances, most utterances are presentations of the objects in a whole sentence; no robot utterances
- wait phase: few user utterances (many of these are short utterances about the object), no robot utterances
- answer phase: few user utterances, many robot utterances
- react phase: many user utterances (however, fewer than in the present phase), most of them praise or comments about wrong robot utterances or presentations in whole sentences and in short utterances

With respect to the expectations of the users it can be concluded that they develop target-based expectations about the robot based on its behavior during the task and not by asking the robot about itself. Such questions were only used very few times. When comparing the two object-teaching studies it became obvious that they were more prominent if the robot often failed to produce a behavior in an appropriate time. Hence, if the users cannot develop expectations about the robot because it does not do anything they could be based on, they actively search for sources of target-based expectations.

Moreover, the users frequently signaled whether their expectations were confirmed (praise) or disconfirmed (comments about wrong robot utterances). Thus, (dis-) confirmation influenced their behavior.

Finally, the results showed that the users' behaviors and the sequence in which they occurred allowed to differentiate success, failure, and confirmation in the object-teaching scenario. This points to the fact that many users behaved similarly in these situations and must also have

perceived them similarly. Hence, the robot behavior was modeled in a way that clearly signaled the state of the interaction to the users and caused similar expectations about what would be an appropriate next behavior.

4.2.4 Analysis of gesture in the object-teaching study

Coding

The analysis of the gestures had its starting point in the coding scheme that has been introduced in Section 4.1. This scheme included the following behaviors:

1. Presenting the object
2. Moving the object once
3. Moving the object continuously
4. Moving the object closer to the robot
5. Manipulating the object
6. Looking at the object
7. Pointing at the object
8. Imitating actions that can be performed with the object
9. Holding the object

When starting to work with the scheme in the contxt of the second object-taching study, it was first checked for compliance with the definition of gesture that had been developed meanwhile (see Section 3.2.2). The term here describes deliberate movements of the arm with sharp onsets and offsets, which are excursions from the previous position. The movements are interpreted as addressed utterances that convey information. In the process of checking whether the gestures comply with this definition, it became obvious that looking at an object (behavior 6) was not a gesture but a gazing behavior because the definition only includes deliberate movements of the arm, the head, and other parts of the body. Therefore, this behavior was taken out of the scheme. However, it motivated the decision to analyze gaze behavior individually (see Section 4.2.5). Moreover, holding the object (behavior 9) was removed because it lacks the intentionality of the movement. A new code was introduced after having a first look at the data: the beat gesture that usually accompanies speech (McNeill, 1992; see Section 3.2.2). In the first study, the beat gestures had either been annotated as continuous movements or were ignored if they were less pronounced. At this point of time their importance was not recognized with respect to the theoretical background and the research questions. Now, they are taken into account because they are not mere continuous movements but have a very specific function in highlighting certain aspects of a phrase (McNeill, 1992). Beat gestures can much more specifically guide attention to a certain part of a sentence as the following example shows. If the user says „This is a cup“ the beat might co-occur with the word „This“ which would highlight the object as a physical entity. It might also co-occur with the word „cup“ which would then highlight the name of the object. The beat might also accompany the whole utterance. All three cases will be analyzed later in this section. When the beat occcurs is important to determine the expectation of the user. If it highlights the object, the users probably expects the main part of the task to be

the visual recognition of the object. If it highlights the name, they expect the more important part to be to teach the word. As a result of the changes the new coding scheme looks as follows:

1. Presenting the object
2. Moving the object once (up, down, to another position, rotate)
3. Moving the object closer to the robot
4. Moving the object continuously (back and forth, up and down, to different positions, rotate back and forth)
5. Pointing at the object
6. Manipulating the object (open the book/bottle)
7. Imitating actions that can be performed with the object (drinking, eating, reading, etc.)
8. Beat

Figure 4-2. Gestures in the object-teaching study
(From left to right: presenting the object [behavior 1], pointing at the object while holding it [behavior 5], pointing at the object while it is on the table [behavior 5], manipulating the object [behavior 6], imitating actions that can be performed with the object [behavior 7])

Results

The coding scheme was analyzed regarding the type of gesture of each behavior (see Table 4-12). Behaviors 1, 4, and 5 are *deictic gestures* (see Rohlfing (to appear) and Section 3.2.2), because they direct attention to a certain object. However, how they do that differs (Clark, 2003) (see Section 3.2.2). For the authors' purposes the presentation of the object (behavior 1) is a way of *placing* the object. The users put the objects in a certain position and then leave it there for a long time so that the robot can see it. Pointing at the object (behavior 5) is a way of *directing* the attention of the robot to the object. The same is true for the continuous movement of the object (behavior 4). For the following analyses, it is assumed that the continuous movement is mainly an action of directing attention to the object because it is moved and thus salient to the robot. Moreover, the phase of maintaining the object in place is very short in this case which is typical for actions of directing attention to objects.

Table 4-12. Overview of gesture types in the object-teaching task

behaviors	gesture type
1, 4, 5	deictic
2, 3	deictic/manipulative
6, 7	manipulative/function
8	beat

This differentiation of the gestures could also be connected to a difference in the meaning of the gestures. Therefore, in the following the gestures in this group will not only be analyzed regarding differences to other groups but also regarding differences within the group.

Moving the object once (behavior 2) and moving the object closer to the robot (behavior 3) can also be viewed as *deictic gestures* but, additionally, they have a *manipulative* component (Rohlfing, to appear). The object is manipulated in a way to facilitate the task for the robot by searching a better place where the robot has better access to the object visually. In this sense, both behaviors can be regarded as one, or moving the object closer to the robot could be regarded as one special case of moving the object once. While this is certainly true, both cases have been coded separately because of the hypothesis that moving the object closer to the robot is connected to the concrete expectation of the user that the object is too far away for the robot to recognize it. Whether actually differences between the two behaviors can be found or if they are used in different situations, will be evaluated in the following. Both behaviors could be merged into one if this is not the case. Manipulating the object (behavior 6) and imitating actions that can be performed with the objects (behavior 7) also have a *manipulative* component. However, they are less deictic but rather *functional* because they explicitly point to the function of the objects. Imitating actions that can be performed with the objects is more concrete in pointing out the function than manipulating the object. Finally, the last group consists of *beat* gestures (behavior 8).

As a first result of this grouping according to type, it is noticeable that *no iconic gestures* were carried out. This is in line with research by Iverson, Longobardi, and Caselli (1999) who found the same for gestural interaction between Italian mothers and their children (see Section 3.2.2). The function of the gestures was to refer to the immediate context and to reinforce the message conveyed in speech to facilitate speech understanding. Accordingly, if the users expected that the robot had the cognitive abilities of a young child rather than of an adult, this would explain why they did not use iconic gestures. However, the explanation might also be much simpler. The situation as such might make the usage of iconic gestures redundant. Iconic gestures would require the movement to be transferred to the object it refers to, but as the participants can show the objects directly because they are available right in front of them, the description of the objects with the help of an iconic gesture is simply not necessary. The gestural behavior would probably change if the objects were not there. This finding points to the influence of the scenario on the usage of gestures and to the fact that the coding scheme is restricted by and to the scenario.

To test the reliability of the coding scheme within this scenario interrater agreement between two raters was calculated for nine files (41%) that were chosen randomly but balanced for the distribution between the conditions (four files from the negative trials and five files from the positive trials). In a first step, the overlap of the annotations of both raters in time was analyzed in order to make sure that they recognized the same movements of the subjects as gestures. The overlap of both annotations was 94%. This agreement is high, especially because the users did not gesture continuously but in 68% of the time. In a second step, Cohen's Kappa was calculated. The Kappa value was .71, which according to Wirtz and Caspar (2002) is a good agreement. However, further analysis will show that presenting the object (behavior 1) occurred

much more often than the other behaviors. This certainly influenced the value. The Kappa value for the types of gestures was also .71 which indicates that the gestures of one type were not confounded with each other but in case of disagreement rather gestures of different types were annotated. For example, in 37 cases one of the coders coded a continuous movement whereas the other coded single movements with pauses, depending on how the gestures were segmented. Moreover, one coder annotated 42 gestures that the other coder did not identify as gestures. These were mainly gestures of presenting the object which the other coder thought were not gestures but rather instances of holding the object without gesturing, or gestures of single movements (17) which the second coder identified as movements that result naturally when holding an object and are not intentional gestures. Thus, the intensity of the gesture seems to play an important role and might be a valuable question for future research.

2147 gestures were annotated with the help of the coding scheme that has been shown to be sufficiently reliable for this scenario. The results of the analysis of the annotations with SALEM are depicted in Table 4-13.

The table shows that the participants most often presented the objects to the robot (behavior 1) (40.33%). This behavior was often interrupted by a beat gesture when the person spoke (15.93%). Moving the object once (behavior 2) was also very common (22.08%). Moving the object closer to the robot (behavior 3) only took place 23 times. Due to the small number of occurrences, the fact that moving the object once and moving the object closer to the robot represent the same movement with different directions, and are very similar with regard to mean duration, a distinction was not found to be useful for the following analyses. Hence, the two behaviors will be grouped.

In 10.25% of the cases, the participants moved the object continuously to attract the robot's attention. This kind of deictic gesture was much more common than pointing at the object (5.73%) in the scenario presented here. Obviously, when the objects are manipulable the participants prefer to manipulate them instead of pointing at them. In fact, only two participants pointed a lot (47 [22.82%] and 48 times [49.48%]), hardly ever lifted up the objects, or tried other behaviors. The person who pointed the most, apart from pointing mainly moved the object once (behavior 2) (19.59%) or moved the object continuously (behavior 4) (15.46%). He moved the objects on the table or lifted them up quickly to then put them back right away. Two users

Table 4-13. Descriptive statistics of the gesture behaviors

code	count	count of all annotations (%)	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)
1	866	**40.33**	0.01	87.75	**6.37**	7.42	4.18
2	474	22.08	0.01	4.19	1.06	0.57	0.91
3	23	1.07	0.30	2.24	1.04	0.52	0.86
4	220	10.25	0.44	16.73	3.31	2.43	2.51
5	123	5.73	0.33	43.02	3.63	5.02	1.89
6	36	1.68	0.49	10.69	3.35	2.57	2.39
7	63	2.93	0.58	21.42	2.86	2.97	2.06
8	342	15.93	0.29	7.24	1.29	0.78	1.15
overall	2147	100	0.10	87.75	3.71	5.50	1.77

did not point at all; three only very infrequently in one of the trials. The users who pointed ten times or less usually did so while holding the object in their hands. Because they pointed so little or only pointed and did not use other behaviors, with pointing alone Clark's (2003) assumption that directing-to actions are shorter than placing-for actions could not be researched here. Therefore, the mean durations of both directing-to actions, pointing (behavior 5) and moving the object continuously (behavior 4) were compared to the mean duration of the placing-for action presenting the object (behavior 1) for all trials. In fact, the mean duration of the placing-for gestures was found to be significantly longer than the mean duration of the directing-to gestures (T-test (two-tailed), df=21, T=3.481, p=.002). Hence, the distinction between these two types of deictic gestures seems to be conclusive here. However, all three gestures belong to the category of deictic gestures. In the following, this gesture type will be compared to the other types. Table 4-14 gives a first overview of the results.

From Table 4-14 it can be seen that the mean durations between the gesture types varied considerably. Hence, the assumption was made that the gesture types can be discriminated based on the mean duration of all trials. Therefore, T-tests were calculated that tested all types against the other types for all participants. Table 4-15 shows the results. Since not all participants performed gestures of the types manipulative/function and beat, the degrees of freedom vary between the tests.

First of all, Table 4-15 shows that the mean durations between all gesture types vary significantly. The only exception is the comparison between deictic/manipulative and beat. The trend was that the beat gestures took longer but this difference did not reach significance. However, the results of all other tests are clear which allows for three assumptions: (a) the grouping of the gesture types is valid, (b) the gesture types can actually be discriminated based on their mean durations, and (c) the different mean durations imply different functions of the gesture types.

Table 4-14. Descriptive statistics of groups of gesture types

type	count	count of all annot. (%)	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)
deictic (1, 4, 5)	1209	56.31	0.01	87.75	**5.53**	6.69	3.36
deictic/manipulative (2, 3)	497	23.15	0.01	4.19	1.06	0.57	0.91
manipulative/function (6, 7)	99	4.61	0.49	21.42	3.04	2.83	2.21
beat (8)	342	15.93	0.29	7.24	1.29	0.78	1.15
overall	2147	100	0.10	87.75	3.71	5.50	1.77

Table 4-15. T-tests (two-tailed) for gesture types

type	df	T-value	significance (p)
deictic - deictic/manipulative	21	6.721	.000
deictic - manipulative/function	11	3.816	.003
deictic - beat	19	5.951	.000
deictic/manipulative - manipulative/function	11	-3.674	.004
deictic/manipulative - beat	19	-1.894	.074
manipulative/function - beat	11	3.044	.011

The deictic gestures took significantly longer than all other types of gestures. It can be suggested that their function is to *maintain* the robots' attention to the objects which is very important with respect to the task. In order to teach the objects it is essential that the robot attends to them carefully as long as is necessary to learn them. Therefore, the participants spent a long time on deictic gestures.

Deictic/manipulative gestures also have a deictic component. However, their mean duration was much shorter. This can be explained by their different function. This type of gestures is also used to steer the robot's attention. However, it is not produced to *maintain* but rather to *attract* attention in the first place. Thus, behaviors of this category were used when the users were no longer sure whether the robot was still attending to the object and they expected that a manipulation makes it easier for the robot to recognize the object. The gesture type can be seen as a preparation for deictic gestures because the attraction of the attention should logically be followed by its maintenance. With the help of the transition matrix this was found to be true in 81.86 % of the cases. With a probability of 14.29% deictic/manipulative gestures were succeeded by a beat gesture. This implies that the attraction of the attention was directly followed by an utterance regarding the object.

To ease the interaction for the robot is also a function of the gestures of the type manipulative/function. However, these differ from the deictic/manipulative gestures in that they do not mainly serve to attract the robot's attention but to explain the function of the object. Since this endeavor requires much more complex concepts about the meaning of the objects, it is comprehensible that these gestures had a longer mean duration. Most commonly they were preceded by deictic gestures (67.68%) and also succeeded by deictic gestures (95.09%).

Finally, the function of beat gestures has been introduced in Section 3.2.2. They serve to stress certain parts of utterances. Hence, in contrast to the other gesture types they rather highlight what is important about the utterance and not what is important about the object. The transition matrix showed that beat gestures were usually introduced by deictic gestures (76.11%) and were also succeeded by these (87.13%). However, as has been mentioned above, they may also be preceded by deictic/manipulative gestures (20.76%). Accordingly, the participants made sure that the robot was attentive before using beat gestures, i.e., before saying the important part of the utterance with respect to the task. For now, it shall be concluded that beat gestures as they co-occur with speech are rather short because utterances in general are shorter than most gestures. Furthermore, beat gestures often accompany only single parts of the utterance. The interrelation between gesture and speech will further be analyzed in Section 4.2.6.

Based on the functions of the gestures that have been described above, one could assume that some of them were more important in the positive or in the negative trials. Several hypotheses can be generated from these findings. Some shall be mentioned here: (a) the participants use more deictic/manipulative gestures in the negative trials because they assume that the robot repeatedly names the objects wrongly because it cannot recognize them and needs help to find them; (b) the participants use more beat gestures in the negative trials to stress their words and to help the robot to recognize the important word (i.e., the object name) in the sentence; or (c) the mean duration of the pointing gestures is longer in the negative trials because the users more strongly attempt to direct the robot's attention to the object. However, none of these hypotheses

could be approved. In contrast, the behavior of each participant regarding gesture types and mean duration of the gestures was found to be very stable. No significant differences between the positive and the negative trials were discovered. Also the amount of gestures per second was highly correlated (p=.000**). Thus, what gestures are used is a question of personal preference but not of how effective the gestures seem to be with regard to the task. This is in line with the finding of the first object-teaching study that the participants do not consciously adapt the gesture behavior. As has been argued in this context, the conscious change in behavior seems to depend on the feedback of the interactor. BIRON did not produce gestures and the participants could not know how the robot perceived their gestures. Therefore, even though the situation changed in general, the situation with respect to the gestures remained the same, and the variation of the success of the task did not influence the users' behavior.

The question remains as to whether the gestures varied in the phases of the interaction. A first analysis revealed that the number of gestures per second differed for the four phases (present 0.73, wait 0.276, answer 0.504, react 0.405 gestures/second). In the present phase the users gestured the most; in the wait phase the least. This is in line with the assumption that gestures accompany speech. The users were also found to speak most in the present and in the react phase, and in these phases they gestured. However, the rate of gestures was also high in the answer phase, i.e., when the robot spoke. This result is not conclusive right away. In the following, it will be analyzed which gestures were used in the phases (see Table 4-16). Maybe, this analysis can explain what gestures occurred in the answer phase and why.

The answer phase was dominated by deictic gestures. What can be seen in the videos is that the participants continued the deictic gestures while the robot spoke and they concentrated on listening to it. It can be hypothesized that this is the case because the deictic gestures cause the lowest cognitive effort for the users while they are listening to the robot.

At first sight, it was very surprising that the participants used least deictics in the present phase. They presented the object to the robot (behavior 1) in 41.12% of the cases which was less than in all other phases. However, when taking a closer look at the results it became clear that in the present phase, which was the first phase of the task, the users produced more gestures of the type deictic/manipulative (19.23%) to obtain the robot's attention than in the other phases. Moreover, the users spoke most in this phase (see Section 4.2.3) and, thus, used the most beat gestures (20.42%).

Table 4-16. Descriptive statistics of gesture types in phases

	present			wait			answer			react		
gesture type	**count**	**%**	**mean duration (sec.)**	**count**	**%**	**mean duration (sec.)**	**count**	**%**	**mean duration (sec.)**	**count**	**%**	**mean duration (sec.)**
deictic	617	56.76	7.40	972	72.00	6.37	544	**84.08**	**8.44**	581	64.99	7.78
deictic/ manipulative	209	**19.23**	1.12	175	12.96	1.06	60	9.27	1.05	125	13.98	1.07
manipulative/ function	39	3.59	3.74	66	**4.89**	3.22	25	3.86	4.77	34	3.80	4.15
beat	222	**20.42**	1.38	137	10.15	1.43	18	2.78	1.09	154	17.23	1.35
overall	1087	100	4.83	1350	100	5.03	647	100	7.49	894	100	5.60

The react phase was rather similar to the present phase as they overlap to a large degree. However, it is characterized by a few more deictic gestures and a few less deictic/manipulative gestures. The reasons for this can be seen in the fact that the attention towards the object had already been established beforehand if the reaction was a presentation.
In the wait phase most of the gestures of the type deictic/manipulative were produced. This was probably due to the fact that the users were waiting and if the feedback of the robot took some time they tried to enrich their first explanation with a gesture. Moreover, in the wait phase the cognitive load was lower and they had the capacities to produce the more complex manipulative/function gestures. That one reason for their usage was actually cognitive load can be seen in the fact that the mean duration of the manipulative/function gestures was shortest in this phase. It seems that the gestures were easier and faster to produce because they were used without other modalities such as speech.

Conclusion

To conclude, in all phases deictic gestures were prominent. However, each phase has a kind of own "gesture profile" which can be attributed to its function. In the present phase the users tried to attain the robot's attention and, therefore, produced deictic/manipulative gestures. Moreover, they spoke most in this phase and the utterances were frequently accompanied by beat gestures. In the wait phase the highest amount of gestures of the type manipulative/function were used to enrich the explanation of the object by its function. The answer phase was characterized by the most deictic gestures. The profile of the react phase is similar to the present phase because of the overlap between both. However, the react phase is characterized by fewer deictic/manipulative and beat gestures and more deictic gestures. Hence, its function is less to attain attention and more to maintain it.
In accordance with these findings and the model, the users chose the gestures based on the situation. How they perceived the situation determined which gestures with their respective functions (attract attention, maintain attention, highlight parts of utterances, explain the function of an object) were produced. In this context, it was found that iconic gestures with their function to depict the object abstractly were not necessary in the object-teaching situation with the objects being present. Conceptually simpler gestures that referred to the immediate context directly were sufficient.
In contrast to the model, robot behavior did not influence the choice of gestures because it was not part of the physical social situation. Thus, the users' expectations could not be disconfirmed directly and their gestural behavior changed unconsciously as was found in the first object-teaching study. This also explains why the users did not adapt to the robot, their behavior did not change between the positive and the negative trial, and interpersonal differences were obvious. Hence, it can be assumed that the influence of the users' personality and personal style of interaction becomes stronger, if the robot does not produce behaviors of a certain modality.

4.2.5 Analysis of gaze in the object-teaching study

Coding

As has been mentioned above, the coding scheme for the gestures at first included the behavior "looking at the object". The revision of the scheme resulted in the idea to code gaze as a modality of its own. Section 3.2.4 has described that gaze was coded in three categories: looking at the robot, looking at the object, looking somewhere else. This section introduces the results of the analysis of the annotations with SALEM.

Results

Altogether, 2 hours, 53 minutes, and 30 seconds of gaze behavior were annotated. Their evaluation showed that the participants looked at the robot 67.86% of the time, 15.83% at the object and another 16.32% somewhere else (either the table with the objects, or somewhere else in the room) (see Table 4-17). As stated in Section 3.2.4, Argyle (1988) has reported an average looking time at the other person in HHI of 60%. Therefore, roughly 68% is rather high in relation to his results because, in general, people engaged in task-related HHI look at each other even less and focus more on the object. Argyle (1988) also reports that one glance at the other person is about three seconds long. Glances at the robot here took a mean of 3.89 seconds. However, the average length of glances was shorter here, too (2.66 seconds), because the glances at the objects (1.25 seconds) and somewhere else other than the robot and the object (2.18 seconds) were rather short.

Regarding the duration of glances, it shall be noted that the longest glance at BIRON took 61.29 seconds. This behavior would probably not be acceptable in HHI with an interaction partner whom someone does not know well. This long glance took place in the wait phase of a negative trial. Therefore, it can be assumed that the user waited for a reaction of the robot which only caused a low cognitive load. This long glance and the high standard deviation also increased the mean of the glances considerably. Therefore, also the median was taken into account (see Table 4-17). The median shows that half the glances at the robot were shorter than 2.01 seconds which is close to half of the mean.

Table 4-17. Descriptive statistics of gaze direction

gaze direction	mean duration (seconds)	minimum duration (seconds)	maximum duration (seconds)	standard deviation duration	median duration (seconds)	duration of all annotations (%)
robot	**3.89**	0.06	61.29	5.19	2.01	**67.86**
object	1.25	0.19	9.02	1.04	0.88	15.83
else	2.18	0.11	15.56	1.91	1.60	16.32
overall	2.66	0.06	61.29	3.87	1.33	100

Table 4-18. T-tests (two-tailed) for the medium duration of glances (robot, object, somewhere else)

Gaze direction 1	T-value	df	significance (p)
duration of glances at robot - duration of glances at object	5.09	21	.000**
duration of glances at robot - duration of glances somewhere else	3.67	21	.001**
duration of glances at objects - duration of glances somewhere else	-7.42	21	.000**

Table 4-19. Descriptive statistics of the gaze directions in the phases

	present			wait			answer			react		
direction	**count**	**%**	**mean duration (seconds)**	**count**	**%**	**mean duration (seconds)**	**count**	**%**	**mean duration (seconds)**	**count**	**%**	**mean duration (seconds)**
robot	748	**62.49**	6.40	1169	58.28	**5.30**	626	56.50	6.10	767	47.58	5.52
object	353	29.49	1.30	568	28.32	1.15	204	18.41	1.50	426	26.43	1.38
else	96	8.02	2.20	269	13.41	2.19	278	25.09	2.47	419	25.99	2.50
overall	1197	100	4.56	2006	100	3.71	1108	100	4.80	1612	100	3.64

Even though the overall time that the participants spent looking at the object and somewhere else was almost identical, the number of gazes towards objects was significantly higher (1319 [34% of total number of glances] vs. 778 [20%]). A comparison of the mean durations of all gaze directions revealed that they differed significantly (Table 4-18).

Thus, the time that a glance takes is determined by where the user looks. The question now is whether it is also influenced by the success of the interaction, i.e., whether it differed in the positive and negative trials. In general, the behavior of the participants was rather stable between the trials. For all three values the mean duration of gazes was highly correlated between the positive and the negative trials (gaze at robot .94, p=.000**; gaze at object .81, p=.003**; gaze somewhere else .752, p=.008**). However, the duration of glances at the robot was found to be significantly longer in the negative trials than in the positive ones (mean durations 4.28 seconds [negative] and 3.37 seconds [positive]; [T-test (two-tailed), df=10, T=2.47, p=.033*]). On the contrary, glances at the objects were significantly longer in the positive trials (mean durations 1.17 [negative] vs. 1.34 [positive]; T-test (two-tailed), df=10, T=2.623, p=.025*). No significant difference in length of glances in directions other than the robot and the object was found (means 2.21 [negative] vs. 2.14 [positive]). These results show that the robot was more closely monitored when the interaction was problematic. The finding is also highlighted by the fact that significantly more gaze shifts were carried out in the positive compared to the negative trials (.41 gaze shifts per second [positive] vs. .35 per second [negative]; T-test (two-tailed), df=10, T=2.406, p=.037*). In the negative trials the users less often looked away from the robot. In the following the gaze behavior will also be compared for the phases. In Section 3.2.4 it was found that listeners' glances are longer than speakers' glances (Argyle, 1988). Table 4-19 shows that this was not the case here. Gazes in the wait phase were shortest (mean duration 5.30 seconds). In this phase, the users did not gesture and speak as much (see Sections 4.2.3 and 4.2.4). They handed the turn back to the robot in the wait phase. In HHI the function of gaze while listening is to pick up nonverbal cues.

The glances of the users in HRI might be shorter because here the nonverbal signals are not as complex as in HHI and in fact, the robot does not yet speak in the wait phase. Overall, the users still look a considerable amount of time (58.28%) at the robot in this phase. This percentage was only a little lower in the answer phase (56.50%) where in turn the glances at the robot were a little longer (mean duration 6.10 seconds). However, they were still shorter than in the present phase. Again, this is not in line with Argyle's (1988) finding that listeners look more and in longer glances. Instead, the participants were found to look away more often. This is due to the

fact that some participants put the objects back on the table while the robot was still speaking. However, the relatively high percentage of looking somewhere else in this phase was also caused by the users frequently looking to the side when the robot said something. Apparently they did so to better concentrate on what it was saying. Consequently, in contrast to HHI, the speech of the robot alone seemed to cause a high cognitive load and looking at it did not improve understanding because it did not underline what was said by facial expressions or lip movements.

However, gaze direction was also influenced by the task as the following result shows. The amount of looking at the robot was lowest in the react phase (47.58%). In this phase the users looked much more at something else than the robot or the object than in the other phases. This finding can be attributed to the fact that the participants used the react phase to choose a new object from the table which was coded as looking somewhere else until they gazed at a concrete object. Thus, the requirements of the task influenced where the users looked. Apart from these findings, the gaze behavior in the phases was rather similar.

Conclusion

Unlike the other modalities, gaze does not allow for a discrimination of the phases. At least the discrimination in this case cannot be unimodal but requires the combination with other modalities. Some multimodal results that further enrich the analysis are presented in the following section.

With respect to the model and the users' expectations it can be concluded that in case of disconfirmation of the expectations, i.e. in the negative trials, the users spent more time looking at the robot and monitored it more closely. This again points to the fact that gaze is not only a signal but also a channel for perception (see Section 3.2.4). Moreover, single glances were very long which is unusual in HHI (Argyle, 1988; see Section 3.2.4). This finding is closely connected to how the users perceived the social situation. They did not expect that their gazing behavior led to high cognitive load or arousal, as it would have in HHI.

4.2.6 Analysis of the interplay of modalities in the object-teaching study

While the analysis of single modalities offers rich findings, the interaction is multimodal and, thus, the interplay of the modalities needs to be taken into consideration. Hence, in this section, the functionalities of SALEM shall be exploited that allow for comparisons between tiers. The coding is similar to the analysis of the single modalities.

Results

A question that has been raised above with regard to speech and gesture is the synchrony between both. One finding in favor of synchrony is that 83.93% of the annotated utterances were accompanied by gestures. The question now is how the gestures relate to the utterances in time.

Table 4-20. Relation between types of utterances and gestures

	gestures									
	overlap %	full extend		begin extend		end extend		included		sum
utterance about:		count	%	count	%	count	%	count	%	count
object	93.30	302	22.76	431	32.48	386	29.09	208	15.67	1327
robot	91.76	21	37.50	12	21.43	15	26.79	8	14.29	56
attention	76.04	13	41.94	10	32.26	6	19.35	2	6.45	31
interaction	51.01	138	41.82	133	40.30	49	14.85	10	3.03	330
overall	83.93	330	21.74	588	38.78	370	24.37	230	15.15	1744

Most gestures that accompanied utterances about the object began before the utterance and ended during the utterance (32.48%). This is due to two typical sequences of gestures that accompanied these utterances. Compared to the other types of utterances, fewer deictic gestures were used with utterances about the objects. The first typical sequence that caused this result was that the participants were still moving the object to a certain place in the field of vision of the robot (deictic/manipulative gesture) while they started to talk and then switched to a purely manipulative gesture. Hence, many gestures started during the utterances which led to an evenly high percentage of gestures that began during the utterance and ended after the utterance (29.09%). A second typical behavior was to first use a deictic gesture and to then interrupt this gesture with a beat gesture to highlight the important part of the utterance. The beat gestures, which co-occurred with utterances about the object, shall be discussed here in some more depth because they are most closely connected to speech. Altogether, 329 beat gestures co-occurred with speech. 89% (294) of these accompanied utterances about the object (behaviors 1-6, no beat gestures during behaviors 7 and 8). The comparison of the single speech behaviors led to some inferences regarding the words that the beat gestures stressed because for each behavior the structure of the utterances is known. The fewest overlap with beat gestures was found when the object was not named. For the description of the object without naming it (behavior 4) the overlap was 8.33% (17.95% of the utterances accompanied by a beat gesture) Deictic words (behavior 6) overlapped with beat gestures in 10.68% of the time (17.39% of the utterances accompanied by a beat gesture). Based on these findings, it can be assumed that the beat gestures are used in other utterances that highlight the object name. However, also naming the object in a short utterance (behavior 2) overlapped with beat gestures in only 12.71% of the time (16.53% of the utterances accompanied by a beat gesture). Moreover, short utterances have the highest degree of beat gestures that begin before the utterance (48.78%). This means that in cases when not only the name of the object but also an article was used (for example, as in “a cup”), the participants stressed the article and not the name of the object. These findings are contrary to the assumption that the beat gestures were mainly used with the name of the object. The question remains as to what exactly beat gestures highlighted.

The overlap with beat gestures was highest if the object was named in a whole sentence (23.79%, [68.37% of all beat gestures]). 37.92% of the utterances were accompanied by a beat gesture even though these sentences were longer. 42.29% of the beat gestures were included in the utterance. In this case it is not clear what part of the utterance was highlighted. In 27.86% of the cases the beat gesture began before the utterance. This points to the assumption that the

deictic word “this” at the beginning of the sentence was highlighted. In another 22.89% of the cases the beat gesture began during the utterance and extended its end. In these cases most probably the object name was highlighted.

These numbers show that beat gestures were used both to stress the name of the object and the deictic words. Analysis on word level would be necessary to determine the exact relation. For now it shall be noted that beat gestures are used to highlight the important parts of sentences but what seems important for the participants differs with the situation and its perception by the user.

Let us now turn to the relation between gestures and utterances about object, robot, attention, and interaction (see Table 4-21). The gestures that co-occurred with these types of utterances mostly began before the utterance and lasted until after the utterances. For the utterances demanding for attention (behavior 14) and commenting on the interaction (behaviors 21-26) this result can be explained by the fact that these utterances are shortest as has been found in Section 4.2.3 (0.92 seconds and 0.81 seconds, respectively). However, this is not true for utterances about the robot. It is notable though that in this case (also in utterances about the interaction) deictic gestures were used to a high degree and these were found to have the longest mean duration (see Section 4.2.4). Most utterances about the robot (72.5%) were comments about its ability to see. Connected to this, the participants used long deictic gestures that the robot could actually *see* the object. Hence, the gestures supported or reinforced the speech. The question that can now be asked is whether this is also true for gaze.

In general, gaze behaviors overlapped with speech in 98.01% of the time. This is because gaze direction was annotated whenever it could be identified. Unlike gestures, the users cannot simply use gaze or not, but as in HHI they look somewhere at all times during the interaction. Therefore, the overlap percentage for gaze with speech was higher than for gestures. However, the overlap between the gaze at the robot and the types of utterances and the overlap of gestures and the types of utterances show an obvious parallel: the users seem to look more at the robot when they also gesture during the utterances. Thus, they look least at the robot when making comments about the interaction (behaviors 21-26, 47.18%). However, in contrast to the overlap percentages of the gestures, the participants did not look most at the robot when they were talking about the objects (78.79%) but when they were making utterances about the robot (87.60%). This outcome is not very surprising because it mainly stresses that the robot in these utterances is the object of interest. This is underlined by the fact that the participants spent much

Table 4-21. Relation between types of utterances and types of gestures

utterance about	**gesture type**								
	deictic		**manipulative/ deictic**		**manipulative/ function**		**beat**		
	count	%	count	%	count	%	count	%	sum
object	737	55.53	235	17.71	61	4.60	294	**22.16**	**1327**
robot	41	**73.21**	6	10.71	3	5.36	6	10.71	56
attention	19	61.29	6	**19.35**	3	9.68	3	9.68	31
interaction	246	**74.55**	28	8.48	11	3.33	45	13.64	330
overall	1043	59.81	275	15.77	78	4.47	348	19.95	1744

Table 4-22. Relation between types of utterances and gaze directions

	gaze							
	at robot		at object		else		overall	
utterance about	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)
object	867	78.79	468	16.69	96	2.97	1431	98.45
robot	41	**87.60**	11	**9.00**	1	0.46	53	97.56
attention	28	77.31	13	18.60	0	0	41	95.91
interaction	307	47.18	165	16.39	237	**33.09**	709	96.66
overall	1068	72.43	604	16.45	316	9.13	1988	98.01

less time looking at the object (9.00%) when talking about the robot compared to the other utterance types (see Table 4-22). In contrast, when talking about the object or asking for attention, the interaction focused more on the object and not only on the robot. Finally, when making comments about the interaction, the participants often looked somewhere else. As for gesture this can be attributed to the fact that comments about the interaction were less important for the outcome of the task because the robot did not perform better when praised or told that it did something wrong. However, another reason for this finding can be seen in the structure of the interaction. While the users praised the robot, they often turned away to look for a new object. This finding is supported by the fact that while most gazes at the robot extended the utterances in all utterance types (utterances about object 42.91%; about robot 65.85%, about attention 53.37%) this was less probable for the utterances about the interaction (37.46%).

All these findings show that gaze behavior differed depending on the content of the utterances. However, some gaze behaviors seemed to correlate with speech but were certainly caused by the gestures. When analyzing the relationship between gaze and gesture, it was found that the users spent most time looking at the robot when presenting the object to the robot (overlap 81.69%) and during beat gestures (overlap 81.55%). It can be assumed that these gestures are not very complex and, therefore, the users do not need to pay close attention to them. This result is underlined by the finding that the users spent a higher amount of time looking at the objects when manipulating them which is a much more complex action (overlap 45.24% compared to only 6.83% during presenting the object to the robot).

Conclusion

It has been found above that the gestures and the gazing behavior depend on the type of utterances. With respect to the model, this implies that they are closely connected to the goals of the users in a certain situation. If the goal is central to the interaction, such as teaching the object which is the main task, the users gesture more and also gaze more at the robot. The analysis of the beat gestures showed that they were explicitly used with the goal of highlighting important parts of object-related phrases.

4.3 Conclusion of the object-teaching studies

To conclude the evaluation of the object-teaching studies, this section discusses the findings both on methodological grounds and with regards to the results.

Concerning methodology it was found that the coding schemes from the first study needed to be adapted to the new study even though the scenario remained the same. The first reason for adaptation was the experience gained about speech and gesture with respect to the research questions which had also changed. For example, regarding speech the importance of utterances about the interaction was discovered and is now represented in the new coding scheme. Moreover, it was recognized that beat gestures have another function than simple movements of the object, and gaze behaviors cannot be subsumed in a coding scheme for gestures. Therefore, an approach was developed to code gaze as a modality of its own.

A second methodological finding was that the SALEM approach made the descriptive analysis of the data much easier, quicker, and also more reliable than manual analysis. Therefore, lots of results could be presented here that would have taken a long time to analyze without the semi-automatic approach. As a consequence, SALEM also allows new research questions to be addressed with respect to the interplay of the modalities.

This methodological background heavily influences the findings about the interaction that have been presented in this section. First of all, it enabled the differentiation of the positive from the negative trials by analyzing the outcomes of the object-teaching sequences. Thus, the study for each participant in fact consisted of two trials that differed with respect to the robot's success in learning objects. Moreover, phases (present, wait, answer, and react) could be identified as another feature of the interaction. The timely structure of these phases was identical in the positive and the negative trials. Therefore, they could be seen to be independent from the outcomes of the trials. As well the trials and the phases allowed to describe and to differentiate the physical situation of the interaction and they were the basis to evaluate whether the physical situation in fact influenced the behavior of the users. The behavior of the users in the current section has been analyzed with respect to the single modalities and their interplay.

Speech was found to be characterized by certain sequences of utterances depending on the course of the interaction (success, failure, and clarification). Moreover, the phases of the interaction (present, wait, answer, and react) could be discriminated with respect to the types of utterances. Thus, the physical situation influenced the users' behavior and, in accordance with Argyle, Furnham, and Graham (1981), it can be claimed that verbal categories vary between situations. The most common verbal behaviors throughout the interaction were to name the objects in a whole sentence or in a short utterance. In the interaction with the autonomous robot, naming the objects in short utterances would be disadvantageous. The robot could not handle these utterances (at least not in the version that was evaluated here), because it needed an indicator like "this is" in the utterance to identify the teaching task. Therefore, either the robot needs to be enabled to understand the utterances, or the expectations of the users need to be changed. They would have to recognize that it is not easier for the robot to understand short utterances. This concrete target-based expectation could not be created in the object-teaching studies because in the first study the robot did not provide sufficient feedback and in the second study it was operated by a wizard that acted based on a script. Furthermore, the participants

were found to frequently comment on the interaction, i.e., they praised the robot or told it that it had done something wrong. These utterances could be very useful for future interaction because they clearly showed whether the users' expectations were (dis-) confirmed and if the task was completed or not. Moreover, they can serve to evaluate the quality of the interaction.

The quality can also be judged by the amount of utterances about the robot. A comparison of both object-teaching studies has shown that the users seemed to ask more about the robot itself and about its abilities if it performed worse. These questions could be used as a chance to provide feedback that enables the user to improve their behavior-outcome expectancies and, thus, the interaction.

One main aspect that impairs the interaction seems to be timing, especially that the robot needs too long to answer at all and, thus, disconfirms the users' expectations. Timing is also crucial with respect to turn-taking. It was often found here that the robot interrupted the user because it simply started talking when it had computed the response. In turn, also the users interrupted the robot because they expected that it had finished but it continued talking.

The second modality analyzed in this section was gesture. With respect to gesture types, it was found that no iconic gestures occurred in the object-teaching scenario. It was assumed that this is due to the fact that the participants can manipulate the objects that they teach to the robot. Thus, the complex iconic gestures that need to create an abstract depiction of the objects were not necessary. Other gesture types were more common because they fulfilled crucial functions. Most often deictic gestures were used. Their function was to maintain the attention of the robot that had been attained with deictic/manipulative gestures. Gestures with a manipulative component were used to ease the interaction for the robot. This was also true for gestures of the type manipulative/function. They served to explain the function of an object. Last but not least, the users produced beat gestures to highlight certain parts of their utterances. All gesture types could be distinguished by their mean duration that was closely connected to their function. Attaining the attention of the robot did not take as long as maintaining it, showing the function of an object was a rather complex action that took longer while highlighting part of an utterance needed to coincide with exactly this part of the utterance and, therefore, only took a short amount of time.

For all gestures, it was found that the behavior of the users was rather stable. The gestures that they used differed more between the participants than between the trials. For example, only two users used pointing at objects that were lying on the table as their main behavior while all other users lifted the objects up. However, as the robot did not use gestures itself and did not provide direct feedback to the gestures of the user, the participants did not have a chance to consciously adapt to the robot and to develop target-based expectations. Thus, the personal style of gesturing that seems to play such an important role, might be less important if the robot provided feedback and the users adapted to it. Nevertheless, also a clear influence of the changes in the situations could be shown. The users' behavior differed significantly between the phases of the interaction. Hence, not only the person but also the situation influences the behavior which is in line with the model.

The third modality that was analyzed is gaze. In this context it was observed that the users spent most of the time looking at the robot. They gazed even more at the robot in the negative trials

which indicates that it was more closely monitored when the interaction was problematic. Even though the objects were in the center of the task, glances at them were found to be very short but frequent. In contrast to HHI, it was discovered that listeners did not look more than speakers. In fact when listening to the robot the users often looked away, either to pick up a new object because they foresaw the rest of the utterance or to better concentrate on what the robot said without being distracted by what it did apart from speaking. Consequently, the speech of the robot seems to cause very low cognitive load on the user when it can be anticipated or it causes very high cognitive load when this is not the case. As the robot does not produce facial expressions and lip movements to improve the comprehensibility of what it says, it is even more understandable that the users look away. In this context, as has been found in Section 2.2.6, the appearance of the robot influences the behavior of the users.

Finally, the interplay of the modalities was analyzed. It turned out that the three modalities were strongly interconnected and altered each other. For example, gestures were found to co-occur more frequently with utterances about the object and the robot than with utterances about the interaction. Also the percentage of gazing at the robot, the object, and somewhere else was found to depend on the gestures. The users looked more often at the robot when the gestures were less complex. In general, the percentage of gazing at the robot was highest when the users also talked about it. Thus, the users communicated their expectations with several modalities at the same time.

To conclude, the robot as actor is a part of the situation and both actors together create sequences of behavior that are influenced by the situation. Hence, the human repertoires of behavior are influenced by the robot's repertoires. The repertoires include behaviors that focus on the task; however, also behaviors to attract and maintain attention and to comment on the interaction are required.

5 HRI data analysis of the home tour

This chapter reports the evaluation of the home tour. It starts with a SALEM of the second iteration (see Section 5.1). However, it goes further in that the analysis of the single tasks takes the focus away from the user and more strongly stresses the interplay between user and robot on the interaction level and the way the robot works on the system level (see Sections 5.2 and 5.3). These questions have been researched with SInA. Hence, in this chapter a second methodological approach is applied. Also visual analysis, questionnaires, interviews, and off-talk are taken into account to analyze the home tour.

Objectives

The home tour study was conducted in order to develop a corpus of HRI in a realistic setting with inexperienced users. The interaction is more complex than object-teaching alone. Primarily, the complexity arises from the facts that the agents move in space and that the scenario includes more tasks. These tasks differ in their function for the interaction. Some tasks, such as greeting the robot, are more social; others are more functional and serve to actually get something done with the robot (for example, teach it something, guide it around). The analysis of the home tour focuses on the evaluation and comparison of these tasks in order to identify their influence on the interaction situation. As mentioned above, another objective of the evaluation is to go beyond the interaction level by integrating it with the system level.

Subjects

The analysis is based on a study that consisted of two iterations and was conducted in the robot apartment. 24 subjects participated in this study. All were German native speakers and interacted with BIRON in German. While in the first iteration some of the subjects were university students (average age 25.6 years; seven female, three male), in the second iteration mainly senior people took part (average age 45.5 years; five female, nine male). Even though the participants received a small reward for participation, their main motivation was to come in contact with the new technology. Therefore, they might have had a bigger interest in technology than the average person. Even though the subjects were very interested, they were, nonetheless, inexperienced regarding interacting with robots (average = 1.4 on a scale of 1 [no experience] to 5 [a lot of experience]). This was taken into account when designing the study.

Procedure

The study was composed of the following sequences (see Appendix G). First of all, the participants were welcomed and introduced to the situation. They answered a questionnaire on demographic data and their experience interacting with robots. Afterwards, the subjects were trained in using the speech recognition system, i.e., they were instructed about the proper placement of the head set microphone and were asked to speak some phrases for habituation. The recognition results were displayed in verbatim on a laptop computer. Afterwards, the participants were guided into the room where the robot was waiting ready for operation. They were assisted during the first contact in order to reduce hesitant behaviors and received a tutorial

script for practice. The script included utterances for all tasks that the participants had to complete. During this initial tutorial session, the experiment leader also instructed the users on how to pull the robot to alleviate cases when it got stuck. After the tutorial session, the training script was handed back and the subjects received instructions for the main task which included:

- guide the robot through the apartment, i.e., from the living room to the dining room via the hall (see Figure 1-3)
- teach the living room and the dining room
- teach the green armchair in the living room and the table in the dining room in the first session; the shelf in the living room and the floor lamp in the dining room in the second session

During this part of the interaction, the experiment leader only intervened when prompted by the participants. After the interaction, the participants were interviewed about the experience in general. Moreover, they answered a second questionnaire which included items on liking the robot, attributions made towards the robot, and usability of the robot. More details on the study design can be found in Lohse, Hanheide, Pitsch, Rohlfing, and Sagerer (2009) and Lohse, Hanheide, Rohlfing, and Sagerer (2009).

5.1 Analysis of the home tour with SALEM

This section presents the analysis of the second home tour iteration with SALEM. It aims to introduce the tasks that are part of the home tour and to describe how they relate to and differ from each other with respect to the modalities (gesture, body orientation, and gaze). Finally, the gesture and gaze behavior of the object-teaching task will be compared to the data from the second object-teaching study that have been presented in Sections 4.2.4 and 4.2.5, respectively. Speech is not analyzed here because the participants in the home tour were trained with this respect. In contrast, body orientation was only relevant in the home tour because the users in the object-teaching study stood behind a table and also the robot did not move. Finally, the robot behavior is not comparable because the robot operated autonomously in the apartment and was controlled by a human wizard in the laboratory. Therefore, the data presented here is an addition to the laboratory data focusing on a more complex situation rather than a case for comparison. Only the second iteration is taken into account because (a) it is assumed that one iteration offers sufficient data, (b) the effort of coding can be reduced, and (c) the sample in the second iteration was more representative.

5.1.1 Analysis of the home tour tasks

As has been mentioned above, the home tour consists of several tasks. The identification of these tasks is the first step of TA (see Section 3.3.2) that later on allows each task to be researched separately. Here the tasks are described statistically in order to identify frequencies, differences and sequences.

Results

In the data of the second home tour iteration, 393 interaction sequences were coded that belong to ten tasks. Table 5-1 shows the descriptive statistics for all tasks. Please note that the minimum, maximum, and mean duration are influenced by the fact that the tasks that are not problem-related are interrupted to different degrees by one of the problem-related tasks. As a result, the durations are short if many interruptions occur. On the other hand, the count is higher. Hence, the overall duration percentage of all tasks is the more reliable criterion for comparison because it shows how much time was actually spent on a task.

The table shows that the tasks are structured in three groups. The first group is *social tasks*. Greeting, intro, and farewell belong to this group. They frame the interaction and are mainly used to manage the robot's attention. On first sight, one could think that greeting and intro could be subsumed into one task. However, they were separated because not every user listened to the intro of the robot, and, more importantly, greetings were also carried out during the interaction to regain the robot's attention while the intro really only occurred after the first greeting.

In contrast to the social tasks, guiding, teaching objects and rooms are *functional tasks* that constitute the main part of the interaction. The functional tasks are mandatory to reach the goal of the interaction. The third group subsumes the *problem-related tasks*. As the name indicates, these are problems that can take place during the interaction. 'Obstacle' characterizes a situation when the robot cannot walk any further because something is in its way. 'Register' describes the state when the robot has lost the person and the user needs to register again by saying hello. The 'reset' task allows the user to reset the robot to a consistent state when it does not react anymore. The 'stop' task is initiated by the robot if the users need to say stop in order to finish a previous task. All these problems, in contrast to mere technical problems, can be solved by the users on the interaction level. The four problem-related tasks could also be seen as part of the

Table 5-1. Descriptive statistics of the home tour tasks

	count	**count of all annotations (%)**	**minimum duration (seconds)**	**maximum duration (seconds)**	**mean duration (seconds)**	**standard deviation duration**	**duration of all anno-tations (%)**
social tasks							
greet	45	11.45	0.99	30.23	8.12	6.92	4.99
intro	5	1.27	19.09	57.50	41.81	18.08	2.86
farewell	18	4.58	2.41	23.15	9.13	5.61	2.24
functional tasks							
guiding	109	27.74	3.28	75.07	22.84	12.75	34.00
teaching object	66	16.79	4.14	61.55	19.50	11.58	17.58
teaching room	40	10.18	3.87	37.36	17.71	9.04	9.68
problem-related tasks							
obstacle	25	6.36	8.35	55.38	22.15	9.96	7.56
register	38	9.67	8.19	55.45	19.19	11.55	9.96
reset	18	4.58	13.26	92.14	28.15	19.39	6.92
stop	29	7.38	3.62	29.98	10.60	6.49	4.20
overall	393	100	0.99	92.14	18.63	12.72	100

other tasks. However, they have been coded separately here for two reasons. Firstly, like the other tasks, they are situations that the user needs to solve using certain behaviors. Secondly, analyzing the problems as tasks allows them to be represented in a transition matrix with all other tasks in order to determine when they occur. Thus, the transition matrix offers additional insights in the course of the interaction. But let us first take a closer look at Table 5-1.

The social tasks took less time as the functional and even as the problem-related tasks. While the introduction of the robot only took place a few times, the greeting task and the farewell task both had very short mean and overall durations. Why this is will be further discussed in Sections 5.2 and 5.3. The table also shows that 28.64% of the time was spent on the resolution of problems. The register task alone took up 9.96% of the time of the interaction because it was the most frequent problem. The transition matrix below will explain why this happened. The reset problem did not occur as often; however, it took the longest time to be repaired. This is due to two reasons: firstly, the robot does not tell the user what is wrong and secondly, the robot replies to the reset command that the user has to wait five seconds before saying hello again to reinitiate the interaction. Thus, some time passes without anything happening.

Altogether, most time was spent on the functional tasks (61.26% of the overall duration). This is positive because the interaction was designed for these tasks. 34.00% of the time was used for guiding. Guiding is probably the most complex task in the home tour from a technical point of view because it requires the user and the environment to be tracked while moving. As the transition matrix shows (see Table 5-2), guiding also caused the most problem-related tasks. The guiding sequences were followed by showing the room, which is the task that should follow it according to the study design, as often as by obstacle and register. The obstacle problem is directly related to the guiding of the robot. Obstacle situations mainly resulted when the participants tried to cross the door between the living room and the narrow hallway. That this problem actually interrupted the guiding of the robot can be shown with the high percentage of guiding tasks that succeeded the obstacle task (.88).

Table 5-2. Successor transition matrix for the home tour tasks

		predecessor								
	task	greet	intro	guide	teach object	teach room	obstacle	register	reset	stop
successor	greet	0.044	0.200	0.119	0.212	0.050	0.000	0.081	**0.222**	0.172
	intro	0.089	0.000	0.000	0.000	0.000	0.000	0.000	0.056	0.000
	guide	**0.422**	**0.600**	0.064	**0.273**	**0.250**	**0.880**	**0.486**	**0.389**	0.172
	teach object	0.156	0.000	0.138	0.136	**0.300**	0.040	**0.270**	0.167	**0.310**
	teach room	0.133	0.000	**0.211**	0.030	0.050	0.000	0.081	0.056	0.103
	obstacle	0.000	0.000	**0.202**	0.000	0.000	0.040	0.000	0.000	0.069
	register	0.000	0.000	**0.211**	0.076	**0.200**	0.000	0.000	0.056	0.034
	reset	0.022	0.200	0.037	0.076	0.025	0.040	0.054	0.056	0.069
	stop	0.133	0.000	0.018	**0.197**	0.125	0.000	0.027	0.000	0.069
	count	45	5	109	66	40	25	38	18	29

Register situations were caused by different problems. The robot lost the users because they moved too fast and too far away from the robot, they turned too quickly, or the light changed from one room to the other. With the help of SInA, the problems that took place during the guiding task were identified in depth (see Section 5.3.1).
On the positive side, it can be noted that the problems were hardly ever directly followed by other or the same problems. This shows that the users were actually able to solve them on the interaction level. The transition matrix also displays the structure of the task that the participants were asked to complete. Greeting and introduction are likely followed by the guiding task (.42 and .60, respectively). As has been mentioned above, guiding is then succeeded by teaching the room (.21) and teaching the room is followed by teaching an object (.30) or by another guiding sequence (.25). Also teaching objects is frequently followed by guiding (.27). The main problems that interrupted the teaching tasks differed depending on what was taught. When being taught rooms, the robot frequently lost the user (.20) because it turned to improve its representation of the room and could not focus on the human the whole time. This case is described in more depth in Section 5.3.2. The main problem that occurred when teaching objects was that a previous task, i.e., following, needed to be completed. Afterwards the teaching was resumed (.31).

Conclusion
To conclude, ten tasks were identified and ordered in three groups. The social tasks frame the interaction and are closely connected to expectations from HHI where they occur very frequently; the functional tasks display the goal of the interaction to guide the robot and to teach it objects and rooms; and the problem-related tasks interrupt the social and functional tasks. In connection to the model, the problem-related tasks are disconfirmations of the users' expectations.
This task description is only the first step of the analysis of the home tour data. In the following, a closer look is taken at the whole of the interaction and the tasks with respect to the modalities gesture, body orientation, and gaze.

5.1.2 Analysis of gestures in the home tour

The analysis of gesture contained two parts: the analysis of gestures used to point out objects and rooms to the robot and gestures that did not have a relation to these functional tasks, i.e., all gestures that were produced during other tasks. These groups of gestures are analyzed separately because they serve different functions.

5.1.2.1 Pointing gestures

Coding
In a first step, a data-driven coding scheme for the object- and room-teaching task was developed. Of course, the theory on gestures that has been presented in Section 3.2.2 heavily influenced what was paid attention to. The scheme looked as follows:

1. movement of **whole arm,** pointing with **finger** at one spot of the object
2. movement of **whole arm**, pointing with **finger** at large parts of the object
3. movement of **whole arm,** pointing with **open hand, palm up** at one spot of the object
4. movement of **whole arm**, pointing with **open hand, palm down** at one spot of the object
5. movement of **whole arm**, pointing with **open hand, palm up** at large parts of the object
6. movement of **forearm**, pointing with **finger** at one spot of the object
7. movement of **forearm**, poiting with **open hand, palm up** at one spot of the object
8. sparse movement of **hand** towards the object (hand is not raised above the hip)
9. **beat movement** during speaking
10. **touch** the object with one hand
11. **move** the object to another location
12. other

This coding scheme does not consider the oblique orientation of the palm as suggested by Kendon (2004) (see Section 3.2.2) because the quality of the videos did not allow for such a detailed differentiation between the angles.
To make sure that the scheme was reliable, interrater agreement was calculated for four users (30.76%). In 97.37% of the cases the raters agreed that there was a gesture. However, the codes only agreed in 57.89% of all cases which led to a Kappa of .4676 which is very low. When looking at the data, it became obvious that most often one rater classified a movement as movement of the whole arm, pointing with a finger at one spot of the object while the other rater thought it was a movement of the forearm, pointing with a finger at one spot of the object. Obviously, the raters could not clearly tell forearm and full arm apart. This points to the fact that even if something was shown with the whole arm it was barely ever fully extended and the differences in the angle between upper arm and upper body are rather small between both behaviors. Hence, the differences between the gestures did not mainly depend on the usage of single joints but on the intensity of the gesture, i.e., the degree to which all joints together were moved. Therefore, the behaviors were grouped and the coding scheme was adapted accordingly:

1. movement of **arm,** pointing with **finger** at one spot of the object
2. movement of **arm**, pointing with **finger** at large parts of the object
3. movement of **arm,** pointing with **open hand, palm up** at one spot of the object
4. movement of **arm**, pointing with **open hand, palm down** at one spot of the object
5. movement of **arm**, pointing with **open hand, palm up** at large parts of the object
6. sparse movement of **hand** towards the object (hand is not raised above the hip)
7. **beat movement** during speaking
8. **touch** the object with one hand
9. **move** the object to another location
10. other

This adaptation increased the agreement of the ratings to 83.78% and Kappa to .6929. This value is acceptable for such a low number of cases (38) because every single difference weighs

heavily on the overall result. Again it has to be kept in mind that the movement with the arm, pointing with one finger at the one spot of the object (behavior 1) occurred much more often than the other behaviors.

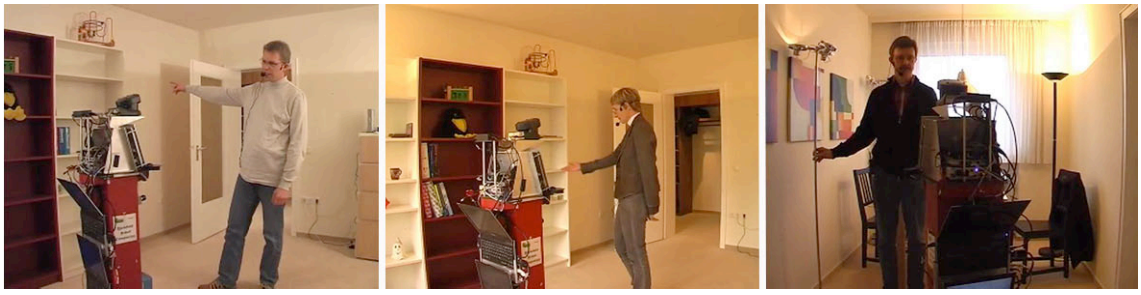


Figure 5-1. Pointing gestures in the home tour
From left to right: movement of arm, pointing with finger at one spot of the object (behavior 1); movement of arm, pointing with open palm at one spot of the object (behavior 3); touch the object with one hand (behavior 8).

Results

Altogether, 113 gestures were annotated for the object-teaching task (1.7 per annotated object-teaching task) and 18 for the room-teaching task (0.45 per room-teaching task). Pointing at rooms was found to be much less common because they surrounded the robot and the human which made pointing gestures redundant.

Similar to the SALEM for the object-teaching study, the gestures were grouped according to their type (see Table 5-3).

Table 5-3. Overview of gesture types in the home tour

behaviors	gesture type
1-6	deictic
8, 9	deictic/manipulative
7	beat

Again no iconic gestures were identified. The gestures will not be analyzed with respect to the types because the large majority of the gestures was deictic (see Table 5-4). As has been mentioned above, most common for the object-teaching tasks were movements of the arm and pointing with the finger at one spot of the object. This result is in line with Kendon's (2004) finding that the index finger extended is used to single out a particular object (see Section 3.2.2). In their data they found that open hand was used if the object was not itself the primary topic of the discourse but was linked to the topic as a location of some activity. Open palm was often found here when the users pointed out a room. Therefore, it seems that open palms are connected to locations which usually differ from objects in that they are more extended in space Very frequently the participants touched the objects (behavior 8). Touch establishes a close connection between the user and the objects in the participation framework. Hence, it can be assumed that touch is used to enable the robot to more easily recognize the object. Therefore, touch is categorized here as a deictic/manipulative gesture that facilitates the interaction.

Table 5-4. Pointing gestures in the object-teaching and room-teaching tasks

gesture	count object-teaching gestures	count of all object-teaching gestures (%)	count room-teaching gestures	count of all room-teaching gestures (%)
1	**65**	57.52	6	33.33
2	4	3.54	0	0.00
3	6	5.31	7	38.89
4	2	1.77	1	5.56
5	2	1.77	0	0.00
6	2	1.77	0	0.00
7	8	7.08	4	22.22
8	21	18.58	0	0.00
9	3	2.65	0	0.00
sum	113	100	18	100

5.1.2.2 Conventionalized and unconventionalized gestures

Also gestures that were no pointing gestures were analyzed. These gestures were either *conventionalized gestures* that have a clear meaning in a culture or *unconventionalized gestures* with no clear meaning.

Results

Only five unconventionalized gestures were coded that covered less than 0.10% of the time of the interaction. Therefore, they will not be discussed further and it shall only be noted here that these gestures do not seem to be important in the home tour. The number of conventionalized gestures was somewhat higher (52). However, these also only covered 0.89% of the time of the interaction. Hence, as a first result it can be concluded that the usage of gestures other than pointing is not common. Nevertheless, the results with regard to conventionalized gestures shall be discussed briefly. These gestures were coded in a data-driven approach, introducing codes for the meaning of the head and arm gestures while coding.

Also for these gestures interrater agreement was calculated for four trials (30.76%). The analysis showed that only 71.79% of the gestures (28 of 39) were recognized as gestures by both coders. The main problem was head gestures of agreement because ten of these were only interpreted as a gesture by one of the coders. When taking a closer look at the gestures it was found that they could usually be interpreted as a movement of the head which is a preparation action before speaking, but also as a short head nod. Thus, intensity of the gestures seems to play a huge role here. Because of this problem, the interrater agreement resulted in a Kappa of .5869. A comparison of all gestures that were coded by both raters reached a Kappa of .9507 (only one different coding). This high number despite the low overall number of gestures that were compared (39 and 28 without the gestures that one of the coders did not recognize) shows that the understanding of the meaning of the gestures was actually shared by the coders. Table 5-5 depicts the gestures that were identified by the rater who coded all data and indicates

whether the gesture was performed with the arms or with the head. The most important gestures are depicted in Figure 5-2.

Conclusion

The results shall not be discussed in more depth here because the number of gestures was so low. Also the analysis with respect to the task does not make sense for this reason. However, it shall be noted that the follow gestures and the stop gestures are of course connected to the guiding task. All other gestures were distributed over all tasks. Further research will have to determine why the participants gestured so little in this scenario. An initial hypothesis is that the users gesture less because the robot does not gesture. Moreover, the robot does not directly display that it has perceived the gestures. Consequently, the users might gesture more if the robot either gestured itself or if it signaled that it perceives the gestures. However, even though the robot had the same abilities with respect to gesture in the home tour and in the object-teaching studies, how the users gestured differed in both cases. These differences will be pointed out in the following section.

Table 5-5. Conventionalized gestures in the home tour

gesture	count
agreement (head) (head nod)	21
doubt (head)	6
follow (arms)	10
follow (head)	2
resignation (arms)	3
stop (arms)	7
surprised (head)	3
sum	52

Figure 5-2. Conventionalized gestures in the home tour
From left to right: doubt (head); follow (arms); stop (arms).

5.1.2.3 Comparison of gestures in the object-teaching studies and in the home tour

The sections about gesture in the home tour and in the object-teaching task have already evoked the impression that the differences between the user's gestures in both scenarios are considerable. In this section they will be compared directly. The comparison can only be made

with respect to the object-teaching task (pointing gestures) in the home tour because this task is the equivalent to the object-teaching task in the laboratory.

First of all, it has to be noted that the users in the laboratory gestured much more than in the apartment (68% vs. 30.85% of the time). This finding has several explanations. First of all, the robot in the home tour studies needed much longer before it replied because it acted autonomously and in the object-teaching study it was found that the users used fewer gestures while they waited for the robot. Another explanation is that the objects in the apartment were considerably bigger than the objects in the laboratory and they could not be manipulated. Certainly the kind of objects plays an important role when people decide how to show them. Even though the users frequently touched the objects in both scenarios, the frequency of touch depended on the objects, with objects that were meant to be manipulated being touched much more often. Nevertheless, it can be concluded that deictic/manipulative gestures occurred in both studies.

Apart from touch, one frequent behavior to single out objects from the environment in the apartment was to point at them with one finger extended (usually the index finger; however, some users used the middle finger instead). Only two people in the object-teaching task used this gesture frequently. However, they also used other deictic gestures instead. Consequently, deictic gestures were found to be the most common in both tasks (object-teaching study 56.31%, home tour study 71.69% of all gestures).

Another commonality with respect to gesture types is that iconic gestures were used in neither study. In Section 4.2.4 it was pointed out that this could be because the users treat the robot like a child or because iconic gestures are not necessary because the objects are right there and their function can be demonstrated with manipulative gestures. These explanations hold for both studies. As iconic gestures can also be used to point out the function of an object, they seem to be even less important in the home tour because there function did not play a role at all. Also gestures of the category manipulative/function were not carried out in the apartment, for example, none of the users turned on the lamp to show the robot how it worked. This difference might be caused by the type of objects or by the instructions to the participants. In fact, the users in the home tour practiced utterances that the robot understood well and utterances about the functions of the objects were not included. As most users kept very much to the utterances they had practiced, teaching the function of the objects did not come to their minds. The usage of utterances that were practiced beforehand might also have caused the users to use fewer beat gestures (7.08% of all gestures in the home tour study vs. 15.93% of all gestures in the object-teaching study) because they merely concentrated on what to say. Moreover, they were explicitly told not to intonate their speech in an exaggerated way.

This comparison shows that the studies agree in the main gesture type that is used. Teaching objects to a robot is clearly dominated by deictic gestures. However, it also points to the differences that the instructions to the participants, and the physical situation (for example, objects that have to be taught, behavior of the robot as an agent) make.

5.1.3 Analysis of body orientation in the home tour

Section 3.2.3 has described the importance of body orientation in the interaction and argued that spatial behaviors in general indicate the structure of the interaction. Therefore, now the results with regard to body orientation of the human and the robot in the home tour are introduced.

Coding

Body orientation was annotated with the coding scheme that has been introduced in Section 3.2.3 (see Appendix A). As explained there, the static and the dynamic body orientation (i.e., movements that change body orientation) were coded separately.

Interrater agreement was calculated for the static orientations in four files (30.76%) and for 195 annotations. The annotations of the static orientations overlapped in 89.08% of the time. This result shows that the annotators agreed most of the time whether the interactors remained in the same body orientation or not. The content of the annotations agreed in 65.13% of the ratings which lead to a Kappa value of only .4245. This value is too low to secure reliability of the coding scheme. Therefore, the disagreements that caused this value were analyzed. It turned out that most disagreements were connected to the differentiation between face-to-face and 45 ^{o} orientations. Thus, it seems that the intervals between the angles were chosen too small. Most often the problem occurred with the 45° right orientation. It was not as common with the 45° orientation to the left side. This points to the fact that the orientations could not be distinguished due to the angle from which the videos were recorded because the camera was positioned most of the time to the right of the robot and the users slightly turned their backs to it when moving right and turned to a face-to-face position with the camera when moving left. This finding was identical for the interrater agreement for the orientation between user and object that was calculated using the same data (agreement 80.95%, Kappa 0.4815). Unfortunately, there are no video recordings from other angles that could prove that the angle of the recordings caused the disagreements. Thus, the possibility has to be taken into account that the coding scheme is too fine-grained and the 90° steps of Kendon's F-Formation that have been applied to HRI by Hüttenrauch and colleagues (2006) (see Section 3.2.3) are more reasonable. Nevertheless, it is still assumed that there is a functional difference between the 45° and 90° orientations as the analysis regarding the tasks will show.

With respect to object-teaching, it can be anticipated that the orientation during the task depends on the object and the way the participation framework is set up. That is why also the orientation of the participants towards the objects was coded with the coding scheme depicted in Figure 5-3.[15]

Results

In the following, first the results of the static orientation will be introduced (see Table 5-6). Table 5-6 shows that the numbers for movements to both sides are very similar. Therefore, the orientations 2 and 6; 3 and 7; and 4 and 8, which correspond to the same angles, will be grouped for further analysis (see Table 5-7). In this second table it becomes even more obvious that the

[15]If the object was large (for example, the shelf) the part of the object that the user pointed at was taken as the point of reference.

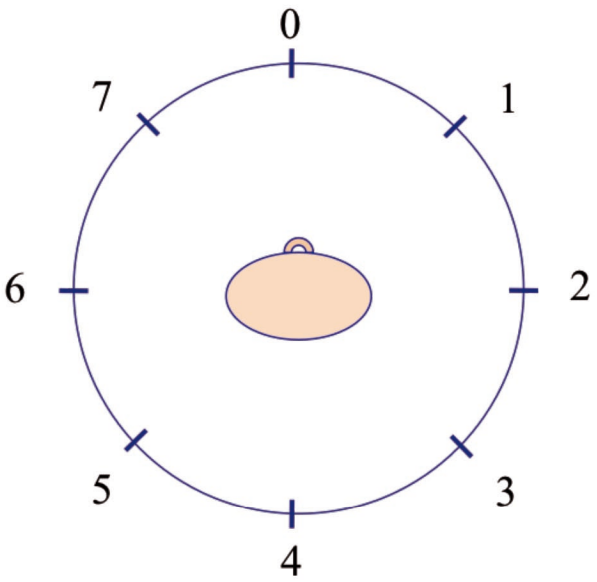


Figure 5-3. Coding scheme for body orientation of the user towards the object

Table 5-6. Descriptive statistics of body orientation in the home tour

body orientation	count	mean duration (seconds)	overlap (%)
1 (face-to-face)	**169**	**31.14**	**73.71**
2 (45° left)	68	10.86	10.34
3 (90° left)	26	5.22	1.90
4 (135° left)	8	2.49	0.28
5 (back to the robot)	12	1.59	0.27
6 (45° right)	60	13.75	11.55
7 (90° right)	22	4.78	1.47
8 (135° right)	4	3.29	0.18
9 (almost side-by-side)	1	5.73	0.08
10 (side-by-side)	1	14.83	0.21
overall	371	19.24	100

Table 5-7. Grouped descriptive statistics for body orientation in the home tour

body orientation	count	mean duration (seconds)	overlap (%)
face-to-face	169	31.14	73.71
45°	128	12.21	21.89
90°	48	5.02	3.37
135°	12	2.76	0.46
back to the robot (180°)	12	1.59	0.27
almost side-by-side	1	5.73	0.08
side-by-side	1	14.83	0.21
overall	371	19.24	100

users preferred orientation with smaller angles. They chose them more often and stayed longer in the orientations. The mean durations of the sequences between face-to-face and 45°; and 45° and 90° differed by a factor of about 2.5. Face-to-face orientations were most common which is in line with the findings of Hüttenrauch et al. (2006).

Hüttenrauch et al. (2006) (see Section 3.2.3) have further shown that the body orientation depends on the task that is solved. That is why the body orientation was analyzed with respect to all home tour tasks (see Table 5-8). It is obvious at first sight that the users chose a very direct orientation in the social tasks. In more than 80% of the time they maintained a vis-à-vis orientation to the robot. In the remaining time they just turned slightly. This is also true for the problem-related tasks with just a few exceptions. Only in the functional tasks did the participants prefer other orientations toward the robot. With respect to the coding scheme, it seems that there is in fact a functional difference between the 45° and 90° orientations. While turning to 45° the users can still signal that they are very attentive. However, in the guiding task the probability of turning to a 90° orientation is bigger because the users prepare to walk away from the robot in order to trigger it to follow.

Table 5-8. Body orientation in the tasks[16]

	body orientation															
	face-to-face		**45°**		**90°**		**135°**		**180°**		**alm. side-by-side**		**side-by-side**		**overall**	
task	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)
social tasks																
greet	36	87.55	8	8.26											44	95.81
intro	6	99.11	1	0.34											7	99.45
farewell	12	80.21	5	16.64											17	96.85
functional tasks																
guiding	107	57.89	66	13.25	35	3.22	6	0.63	9	0.56	1	0.19	1	0.52	225	76.25
object	36	46.57	30	39.55	7	7.42	2	0.53							75	94.07
room	35	72.43	11	7.16	3	0.52	1	0.16	2	0.59					52	80.86
problem-related tasks																
obstacle	32	60.87	7	3.55	1	0.44									40	64.86
register	26	60.37	21	25.49	5	2.48	1	0.36			1	0.13	1	0.26	55	89.09
reset	21	63.40	14	29.04											35	92.44
stop	16	77.71	7	14.16			1	1.40	1	0.36					25	93.63

[16]If the sum of the annotated percent does not reach 100%, this means that part of the sequences could not be coded because the orientation could not be observed in the video. The percentage for the obstacle task is lower than the others because the orientation was not annotated for the time when the users pulled the robot. However, the relation between the annotated cases can nevertheless be regarded as reliable because there is no reason that the users should do something different once the camera is not catching them for a couple of seconds.

As for the coding of the body orientation of the human towards the robot, the angles of the code 1 and 7; 2 and 6; and 3 and 5 agree. Whether one or the other was chosen was determined by the fact on which side the object was located. Obviously, here more objects were located to the right of the users, which was true for both the shelf and the floor lamp that most subjects showed to the robot. The distribution of the annotations for these two objects was very similar. To facilitate the overview, Table 5-10 shows the orientations ignoring the side of the object. The table illustrates that the object was most commonly to the side of the participant's upper body and, as found in Section 5.1.2.1, they point it out with a gesture. The object was nearly as often slightly in the back of the user as it was slightly in front of them. This finding indicates that the users did not necessarily navigate the robot until it was in a position where the object was between them. Figure 5-4 shows the most common constellations of human, robot and object.

Table 5-9. Orientations of the users towards the objects.

orientation	count
0	1
1	10
2	39
3	5
4	0
5	8
6	13
7	5
sum	81

Table 5-10. Grouped orientations of the users towards the objects

orientation of the object	count
between user and robot (0)	1
slightly in front of the user (1, 7)	15
to the side of the user (2, 6)	**52**
slightly behind the user (3, 5)	13
behind the user (4)	0
sum	81

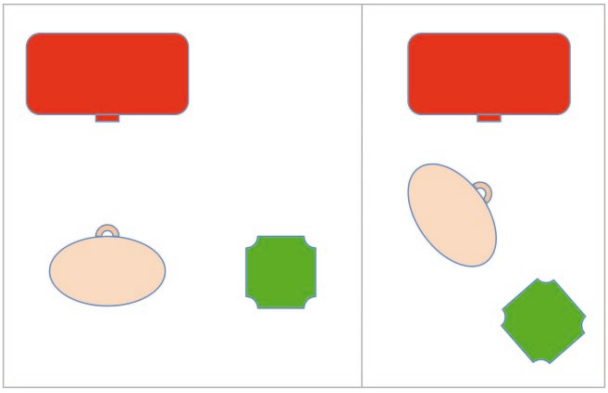


Figure 5-4. Typical participation frameworks of user, robot, and object in the home tour

Next to the static orientations also changes between different orientations were analyzed. Altogether, 262 movements with a duration of 9 minutes, 36 seconds were coded. The mean duration of the position changes was much shorter than the mean duration of the time that the users spent in a certain position (2.20 seconds vs. 19.11 seconds). This is underlined by the finding that the users almost always chose the shortest ways between the orientations. But not only the users changed their orientation; the robot also initiated 96 movements that led to a change. This number of changes was significantly lower over all tasks than the number of movements of the user (T-test (two-tailed), df=10, T=2.464, p=.036*). The robot's movements took 7 minutes, 23 seconds altogether. With a mean duration of 4.62 seconds the movements of the robot were much slower than the movements of the human.

The number of position switches in each task corresponds with the number of static orientations. The users changed their orientation during the guiding task most often by far (3.66 movements/minute). In contrast, the robot most often changed its position in the room-teaching task (1.78 movements/minute), which is in line with the designed robot behavior that makes the robot turn around in order to improve its representation of the room. The task with the second most changes of user (1.97 movements/minute) and robot (1.48 movements/minute) was the register task. This shows that the users tried to position themselves in front of the robot during the task so that it could perceive them well and the robot turned to the user for the same reason. On the other hand, instead of positioning themselves directly in front of the robot, some users walked to the side of it to see whether it followed, which it did as the number of orientation changes of the robot shows. The users did the same during the reset task (2.01 movements/minute). This task was different compared to the register task because the robot did not change its orientation towards the user.

Comparing the movements per minute, the social tasks were found to be rather static (greeting 0.66, introduction 0.57, and farewell 1.10 changes of orientation/minute). The farewell task was accompanied by slightly more movements which is probably due to the fact that the users started to turn away before the robot had replied to the farewell command.

One reason for analyzing body orientation here was the theoretical assumption that it is of importance for the structure of the interaction (see Section 3.2.3). In HHI it is evident that more orientation switches occur at the beginnings and ends of phases of the encounter. That is why the data was analyzed for switches in body orientation at transition points between tasks and at the beginnings and ends of utterances. For the transition points between tasks it was found that fewer switches in body orientation took place than during the tasks. Most changes of body orientation (210, 83.33%) were included in the task and did not extend them. Therefore, this analysis did not confirm the assumption. When having a closer look at the data, it was found that the users switched their body orientation very frequently at the beginning or the end of their utterances (48, 64.0% of changes of body orientation that overlapped with human speech). However, this number also shows that only 29.76% of the total changes of body orientation co-occurred with speech. Hence, these numbers do not clearly point to the structural influence of body orientation on the interaction. This might have three reasons: (a) the body orientation is not used with this purpose in HRI; (b) the results are not conclusive because the scenario is mobile. Thus, the communicative orientation switches are confounded with movements that

have a functional and not a communicative purpose (for example, during the guiding of the robot); (c) the body orientation structures other entities than tasks and utterances.

Conclusion

The tasks could be differentiated based on the orientation of the users. The users most often changed their orientation in the guiding task which was also due to the fact that the robot was moving. Only in this task was it at times found that the users turned their backs to the robot. But as has been mentioned above, the time they spent in this orientation was very short compared to the time they spent in more direct orientations. The object-teaching task was the only task during which the participants turned as many times to 45° and 90° (both summed up) as to a vis-à-vis orientation and also spent the same amount of time in these orientations. This finding underlines that the participants turned when teaching objects to establish a participation framework that enables the robot to optimally perceive the objects which are usually located on the side of the participants. This was not found for teaching rooms where the users most of the time faced the robot. Therefore, the actions of teaching rooms and objects seem to follow distinct rules and need to be differentiated.

Altogether, the static and dynamic body orientation have been shown to differ between the tasks and can, therefore, be used to distinguish tasks from each other and to improve the SA of the robot. However, it could not be conclusively determined how the modality might structure the interaction. This question needs to be evaluated in a scenario that is more adequate in distinguishing communicative switches of body orientation from functional switches that are related to the tasks.

5.1.4 Analysis of gaze in the home tour

Coding

The annotation of gaze behavior followed the rules that were established for the object-teaching task. Accordingly, three behaviors were differentiated: looking at the robot, looking at the object, looking somewhere else.

Results

Altogether, 2 hours, 13 minutes, and 20 seconds of gaze were analyzed. In 90.7% of the time (491 times) the participants looked at the robot, in 2.7% (159 times) at the objects, and in 6.5% (264 times) somewhere else (see Table 5-11).

Table 5-11. Descriptive statistics of gaze direction

direction	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)	duration (%)
robot	0.11	259.53	**14.78**	21.58	8.02	**90.7**
object	0.38	20.72	1.37	1.76	1.05	2.7
else	0.18	17.74	1.98	1.87	1.35	6.5
overall	0.11	259.53	8.69	17.08	2.47	100

Table 5-12. Relation between gaze direction and tasks

	gaze direction							
	at robot		at object		else		overall	
task	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)
social tasks								
greet	50	94.91	1	1.38	7	2.14	58	98.43
intro	14	90.86	0	0	9	9.13	23	100
farewell	20	94.81	0	0	4	4.48	24	99.29
functional tasks								
guide	211	82.07	19	0.90	105	8.28	335	91.25
object	153	87.81	114	**9.85**	18	2.34	285	100
room	60	92.40	12	1.10	26	5.27	98	98.77
problem-related tasks								
obstacle	42	64.19	0	0	8	1.78	50	65.96
register	58	91.19	2	0.17	27	6.41	87	97.77
reset	33	93.03	1	0.08	16	5.45	50	98.56
stop	30	86.41	4	4.87	7	5.71	41	96.99
overall	459	85.75	142	2.44	222	5.59	823	93.79

The overall looking time at the robot and the mean duration of glances (14.78 seconds) were much longer than in HHI (Argyle, 1988, see Section 3.2.4). Even though the mean duration was strongly influenced by some very long glances, the median was also very high (8.02 seconds). The durations of glances at the objects and at places other than object and robots were much shorter (mean 1.37 and 1.98 seconds, respectively; median 1.05 and 1.35 seconds).

The question now is why the users looked at the robot for such a long time. The most probable explanation is that they were waiting for feedback which took longer in this study because the robot was acting autonomously. Actually, the longest glances were found in situations when the users were waiting for the robot to reply. This is in line with Green's (2009) findings that gaze at the robot signals that the users need some feedback (see Section 3.2.4). Moreover, after the interaction many participants stated that they looked at the screen when the robot did not respond (see Section 5.4). Since a screen is something one is supposed to look at and not have to look away from once in a while in order to not offend the interaction partner, this is the most probable explanation. Unfortunately, from the video data that was recorded, it cannot be determined whether the participants actually looked at the screen or at some other part of the robot. An analysis of gazing behavior in the tasks was conducted to shed some more light on this question (see Table 5-12).

The first observation in the table is that the number of glances and the percentage of looking at the objects were by far highest in the object-teaching task. This is in line with the design of the study and the task. The users most often looked somewhere else other than at the robot or the objects during the intro and the guiding task. During the guiding task the participants certainly looked where they wanted to go to. During the intro the frequent glances away from the robot are more surprising because the robot talks about its usage and often refers to pictures on its

screen. When having a second look at the data, it was found that in fact only one user looked away from the robot six times. She turned to the side because the robot had called her by a wrong name (‘Igor’) and she kept laughing about this mistake. Hence, the common gazing direction during the intro really seems to be the robot. But this case also shows that the users might crucially change their behavior if some irregularities occur in the interaction. With respect to the percentage of the time that the users in general looked at the robot, no significant differences were found.

Finally, the gaze direction corresponded with certain body orientations in the tasks. In the social tasks and the problem-related tasks the body orientation was very direct and the users looked at the robot most of the time. In the guiding task the body orientation as well as the gazing direction changed most often. The users spent most of the time looking somewhere else other than the robot and the object and in indirect orientations. The user behavior differed considerably between the two teaching tasks. During object-teaching the users often slightly turned away from the robot and then back towards it as they gazed at the object and then back at the robot. In contrast, during the room-teaching task they gazed at the robot and kept a direct orientation for most of the time. Accordingly, the robot could identify that something was taught based on the utterance and, furthermore, receive information about what was shown via the gaze direction and the body orientation.

5.1.4.4 Comparison of gaze in the object-teaching studies and in the home tour

As stated before, it is assumed here that the repertoires of user behaviors are strongly bound to the situation. In the following, the gaze behavior in the object-teaching study is compared to the behavior during the object-teaching task in the home tour study in order to determine whether it differed in the two situations.

At first sight it can be seen that the users spent considerably more time looking at the robot in the apartment than in the laboratory (88% vs. 68%). Consequently, the percentage of gazes at the objects (9.85% vs. 16%) and somewhere else (2.34% vs. 16%) was much smaller. These differences can be explained on methodological as well as on situational grounds. To begin with the reasons that lie within the situation, the robot in the laboratory was not mobile and therefore it can be assumed that the users in that situation did not monitor it as closely as in the apartment. Another crucial difference is that the screen was turned off in the laboratory while it displayed the Mindi in the apartment. The participants confirmed in the interviews that they often looked at the display for information (see Section 5.4). Moreover, the type of objects might make a difference. In the laboratory, the objects were manipulable which caused the users to hold them in their hands and do something with them. These activities certainly require more attention towards the objects than showing larger objects in a room that in only three cases (2.65% of all cases) were moved. Moreover, picking out the objects in the apartment was much easier due to their size and because the participants were told what specific objects they had to show. This explains why the duration of glances somewhere else other than at the robot and the object was much shorter in the home tour.

However, there are also some restrictions on the comparison from a methodological point of view. First of all, the number of glances that were compared differed considerably (laboratory

7828; apartment 285). Most importantly, the quality of the video recordings and, thus, the granularity of the annotations vary between the studies. While it was easy to get very good recordings in the laboratory because the participants did not move, it was much harder in the dynamic task in the apartment. Therefore, short gazes away from the robot as a result of cognitive load that might easily be recognized in the laboratory might have been overseen in the recordings from the apartment. This assumption is underlined by the finding that the mean duration of glances at the robot was much longer in the apartment (16.06 vs. 3.89 seconds). In contrast, the mean duration of glances at the objects (1.36 vs. 1.25 seconds) and somewhere else (2.10 vs. 2.18 seconds) were very similar in both studies. Therefore, the difference between the two studies mainly results from the long glances at the robot.
This section shows that the comparability between the two studies is limited. Nevertheless, it was found that in both cases most of the time was spent looking at the robot. Moreover, it can be concluded that the gaze direction was in fact influenced by the tasks.

5.1.5 Conclusions of the SALEM of the home tour

In the previous sections, the home tour tasks were analyzed with the help of SALEM with respect to the modalities gesture, eye gaze, and body orientation. Coding schemes were promoted for these modalities. However, with respect to body orientation some uncertainties remain. Moreover, SALEM has been applied to this second corpus of data which underlines its generalizability. It was shown that data could also be compared between the studies.
It was found that the home tour includes social tasks (greet, intro, farewell), functional tasks (teaching rooms and objects, guiding the robot), and problem-related tasks (obstacle, register, reset, stop). With respect to the theory, the problem-related tasks are disconfirmations of the users' expectations because they lead to unexpected situations. SALEM showed that most time was spent on the functional tasks and problems related to them. Accordingly, the home tour is task-oriented. Within the tasks, the users displayed certain behaviors with respect to the modalities gesture, body orientation, and gaze direction. Two groups of gestures were identified: pointing gestures in the teaching tasks and (un-) conventionalized gestures used in the other tasks. Pointing gestures were found to be used much more often in the object-teaching task than in the room-teaching task. This points to the users' expectation that objects needed to be pointed out more precisely than rooms. When analyzing the gestures, it was observed that the coders could not clearly differentiate whether the gestures were performed with the forearm or the whole arm. Therefore, it was concluded that the joints involved in the movement are of less importance for the description of the gestures than the intensity. The coding scheme was adapted accordingly. The most common gesture was pointing with the arm and one finger extended at objects. In contrast, pointing gestures at rooms were more often performed with open hand. Altogether, mostly deictic gestures were identified. This finding is in line with the results of the object-teaching study. Another commonality was that in both studies no iconic gestures were produced because they were not necessary in the situation to reach the users' goals. However, the gestures differed in that more manipulative and beat gestures were performed in the laboratory which was attributed to the differences between the objects and to the fact that the participants in the apartment were trained with respect to speech and while

concentrating on what to say did not use beat gestures. Unconventionalized gestures are movements that are not directly connected to a meaning in a certain culture. They only occurred very few times. Also conventionalized gestures that can directly be associated with a certain meaning were used sparsely. Their usage strongly depended on the specific behavior of each participant. While some used these gestures frequently, others did not use them at all. As has been argued above, this finding supports the assumption that users form target-based expectations only when the robot communicates using a certain modality. As a consequence, the findings for body orientation and gaze, both of which the robot employs, pointed to an obvious impact of the situation. If this is not the case, as was found here for gestures in the positive and the negative trials, the influence of the users' personalities and personal style of interaction is stronger than the influence of the situation.

The analysis of body orientation clearly showed that the behaviors differed between the tasks, which allows to distinguish them from each other, for example, a lot of movement and indirect orientations indicate a guiding task while a direct orientation is more common in social and problem-related tasks. This is a positive first result; however, it has to be interpreted with caution because there were some uncertainties about the coding scheme and it needs to be verified whether changes of 45° can be reliably coded with a more adequate video recording. Furthermore, it could not be conclusively determined how the modality structures the interaction. This question needs to be evaluated in a scenario that is more appropriate to distinguish communicative switches of body orientation from functional switches that are related to the tasks.

The results with respect to gaze behavior were more conclusive. The users most often looked at the robot in the social and the problem-related tasks. It was found that the glances at the robot were very long, longer than in HHI and also longer than in the object-teaching study. This was attributed to the fact that the users sometimes had to wait a long time for the robot to react and gazed at it during this time. Most probably, they also spent a long time gazing at the screen. Hence, the modalities of the robot influence the behavior of the user.

5.2 Analysis of the social tasks of the home tour

Analyzing the data with SALEM has already led to a lot of insights about the home tour scenario. In the following two sections, SInA will add to these insights by taking the focus away from the user and more strongly stressing the interplay between user and robot on the interaction level and the way the robot works on the system level. Thus, concrete problems that occurred during the interaction will be identified.

The tasks of the home tour have been identified in Section 5.1.1. The first group of tasks was social tasks, which will be analyzed in this section. The intro is not discussed further for two reasons. Firstly, the intros mainly consisted of the robot talking about itself and were not very interactive. Hence, they were not very promising in bringing interactive deviation patterns to light. Secondly, only a few intro tasks were annotated. In turn, this section contains a description of how people try to maintain the attention of the robot. This description was not acquired with the help of SInA because it does not follow a prototypical script. However, next to

attaining the robot's attention by greeting and interrupting the attention by saying goodbye, maintaining the attention during the interaction is a central issue.
The functional tasks are elaborated on in the next section. Moreover, problem-related tasks have been identified. These are not discussed in their own chapter. Rather they are taken into account as deviations that took place during the social and the functional tasks.

5.2.1 Greeting the robot (Systemic Interaction Analysis)

Greeting is one of the most common scripts established in HHI. Everybody greets other people many times a day in different situations. Kendon (1990) and colleagues have analyzed greeting sequences of two situations, a wedding and a birthday party, both of which took place in the USA. In the data they identified phases of greeting sequences. The first phase is "sighting", meaning that the people perceive each other (pre-interactional step) and show some distance salutation. In this phase, one person decides whether a greeting interaction should be initiated. If so, the person approaches the other person (depending on location/situation also both may approach each other). While approaching often some kind of "grooming" is evident such as straightening of clothes. Thereafter, the close salutation phase starts often with a greeting ritual such as a handshake or an embrace (close salutations also occur without bodily contact). Both interactants come to a halt during this phase. The length of the halt, however, may vary considerably and it may begin before or after the ritual. Thereafter, the interactants enter the "How Are You" phase exchanging basic information. This phase is often highly formalized. Afterwards, the participants usually move away from the spot where the halt took place and change their bodily orientation to each other. Of course, in HHI there is a lot of variation to this script depending on the situation, the relationship of the people, and the cultural background. The script is more restricted in the interaction with BIRON; nevertheless, it models the most important steps that were also described by Kendon (1990). The *prototypical interaction script* for the greeting task consists of the following steps:

- user walks up to the robot (il[17])
- robot perceives user (sl[18])
- user receives an ID (sl)
- PTA[19] changes to state listen (sl)
- user says greeting ("hallo", "hallo Biron", Biron hallo" ["hello", "hello Biron", "Biron hello"]) (il)
- robot processes utterance (sl)
- Mindi changes to thinking/processing (il)
- robot understands greeting (sl)
- PTA changes to person (sl)
- Mindi changes to standing/userlook (il)
- robot answers ("hallo" ["hello"]) (il)

[17]Interaction level
[18]System level
[19]The PTA is the component for person tracking and attention of the robot BIRON (see Lang et al., 2003).

As was found with SALEM, the users did not only greet at the beginning of the interaction. Hence the script might change during the interaction in that the users do not have to walk up to the robot or that the robot does not need to register them in the system because they are already registered. Only when greeting the robot for the first time at the beginning of the interaction, is the script extended by a kind of "How Are You" phase in which both interactants introduce each other and the robot offers to explain something about its operation.

Before first approaching the robot, some participants' off-talk utterances showed doubts about whether the robot would follow such a human-like script. Some utterances shall here be reproduced to depict these concerns. One user asked "Can I just talk to him?" and another one wondered whether the robot would reply to verbal utterances. Yet another participant asked if she needed to talk normally or to intonate in a certain way. Finally, someone wanted to know if she could solely use the commands that she had practiced before the interaction. Accordingly, the users wondered whether they could talk to the robot, how they should talk to it, and what they should say. The HHI-like greeting script should help to overcome these doubts. However, some *deviation patterns* occurred during the greeting task that will be analyzed in the following.

Altogether, 34 greeting sequences were evaluated. 23 (68%) of these complied with the prototypical script. A comparison with the other tasks will show that this number is rather high. This is probably due to the fact that greeting is an action each of us has practiced many times. As mentioned above, it always follows a similar script which the prototypical robot behavior matches. The number of utterances one would use is restricted. Hence, the users have clear expectations of what are appropriate utterances to the robot and what are appropriate replies.

However, eleven deviating cases were revealed. All of them could be explained with deviation patterns. Table 5-13 depicts all patterns according to what the users do, what happens within the system, what the robot does, the number of occurrences, and the influence on the interaction. In general, deviation patterns should include more cases to be called 'patterns'. However, the deviations introduced in the table were also identified in other tasks, which increases the number of times they occurred.

In five situations, deviations resulted from lacking person perception. This equals almost half the cases that were identified. As pointed out in the prototypical interaction script, the first step of the greeting action is usually that the person steps up to the robot. Thus, in many cases, the

Table 5-13. Deviation patterns in the greeting task

Pattern	User (il)	Robot (sl)	Robot (il)	Influence on HRI	#
Speech understanding					
Errors in speech recognition	utters a greeting	input cannot be interpreted at all	asks user to repeat	users repeat command	4
User speaks while robot speaks	utters a greeting while the robot speaks	partial speech recognition because the robot cannot listen while it talks	no reaction	users wait, after some time repeat the command	1
Repetition of commands	utters a greeting	processing/component is too slow	no reaction	users repeat command, robot answers twice	1
Person perception					
User not (yet) perceived	utters a greeting	no user perceived according to person model	does not react	users wait, repeat command	5

participants greeted the robot while still walking towards it. However, the robot first needs to identify the users visually, to give them an ID, and to switch to the "listen" state. These operations take some time. This was not communicated to the users. Hence, they expected that the robot perceived them right away and was able to react. To improve this situation, it could be communicated to the users that the robot better understands the greeting if they first position themselves in front of it and only then speak. However, this is not in line with user-centered HHI where the robot should adapt to the user.

Four deviations resulted from errors in speech understanding. Three times the robot misunderstood the user and asked her to repeat the utterance. In one case the person spoke while the robot spoke, because the participants did not know that the robot was not listening when talking and expected to get the turn as soon as they said something. In fact, the robot reduces its audio sensitivity in order to avoid echo effects when talking itself. It shall be noted that this is without doubt a technical constraint.

One greeting sequence was in accordance with the prototypical interaction script; however, the robot's reaction was too slow and the user repeated the utterances just before the robot answered. In this case the robot also answered twice to the command which often irritated the users.

All deviations were followed by a prototypical sequence. Either the user was prompted to repeat the utterance or repeated it after a pause. Thus, deviations in the greeting task can easily be repaired and most importantly, the users know how to repair them because of the clear script of the task. This also explains why the mean duration of the greeting tasks was very short compared to other tasks (see Section 5.1.1).

5.2.2 Maintaining the attention of the robot (Visual Analysis)

Once the users had greeted the robot, they needed to make sure they maintained its attention because attention is the basis for interaction (Mills and Healey, 2006). Certainly, attention can be attracted with various modalities, for example, someone waves to signal "Here I am" or someone stands up to give a talk. This section will analyze how the users tried to maintain the attention of the robot in the home tour. The focus is on the particular situation that has been presented in Lohse (2009).

The situation was characterized by BIRON having an incomplete percept of the user. The percept consists of a voice and the visual features legs and face. If the person perception is unreliable, for example, because lightening conditions are poor or because the person has not said something in a while, the robot is not able to match the percept to one person. This state is communicated on the display with a picture of the screen character Mindi (Figure 5-5).

Figure 5-5. Mindi pictures in the situation of poor person perception
From left to right: (A) Poor person perception; (B) Processing; (C) Robot is facing user

Due to the state, the robot does not react to the users until they have done something to improve the perception. In general, solutions to this problem include two strategies: verbal behavior or movement of the user in front of the robot. This state is not part of the explicit dialog but is only communicated using the Mindi display. The situation perception of the users and their expectations determine how they react to the situation. The situation was chosen because it was an unexpected event in the interaction. In fact, it constituted a change in the interaction that should also lead the users to adapt their expectations to the new situation. The modified expectation should then result in a change in user behavior.

The analysis was conducted on the video data acquired in the second home tour study. 26 sequences of the situation described above were identified. In the trial of two users (out of 14) no such sequences occurred. All other recordings contained one to three sequences. In the most basic form, the sequences were characterized by the following actions of the robot on the interaction level. First, the Mindi display changed to "poor person perception" (Figure 5-5, A), the person reacted to the display, the display changed to the "processing" Mindi (Figure 5-5, B), then to the "robot is facing user" Mindi (Figure 5-5, C), and finally the robot said "hello". The average length of these sequences was 13.8 seconds (minimum four seconds, maximum 45 seconds; measured from 'poor person perception' Mindi appearing to Mindi changing to 'processing').

For the evaluation, the events were visualized on a time scale (see Section 3.4). Visualization facilitated the analysis because it allowed for a direct qualitative comparison of the sequences. For this restricted number of plots, it was quite easy to see differences and similarities.

In the analysis, it was found that the new situation changed the expectations of the users. However, there were differences in how they conceptualized the change. All analyzed sequences shared that the users searched a position in front of the robot before starting another behavior to resolve the situation. Users did not only aim to stand in front of the robot body but also tried to be faced by the camera on top of the robot. All users kept walking around and leaning to the side until the robot faced them. Therefore, the 'poor person perception' display triggered the same expectation in all users that the robot could perceive them better when the camera was turned to them.

Another behavior that all participants had in common was that the time they waited before the next action (verbal or movement) stayed the same within subjects. However, it strongly differed between subjects. Some only waited two seconds while others waited ten seconds or more for BIRON to give a feedback. The expected feedback could be verbal but also appear on the screen. The observations clearly showed that when the Mindi display changed to 'processing', the users did not take any more actions and waited for more robot feedback.

Another expectation that was shared by all participants was that some verbal utterance was necessary to regain the robot's attention. However, three different strategies to resolve the situation could be identified that are closely connected to the users' expectations:

1) verbal behavior only
2) verbal behavior first with movement added
3) movement first with verbal behavior added

Strategy 1 was characterized by verbal behavior only. Out of the 26 analyzed sequences, 13 were identified with this group. They mainly took place in short sequences in which it sufficed that the users said something once (either “hello” or something that starts another action like “follow me”) (see Figure 5-6). The average length of these sequences was 7.2 seconds (minimum four seconds, maximum 16.5 seconds). The average time people waited after the last action was 6.2 seconds (minimum 3.25 seconds, maximum 7.5 seconds). Measured from the appearance of the Mindi it was four seconds (minimum 1.5 seconds, maximum 6 seconds). All these numbers do not include the two sequences of one particular user because she took significantly longer than all other subjects and was obviously distracted by some event that was not connected to the interaction.

In addition to the sequences which were resolved after giving a verbal command once, this group also includes two situations (two different subjects), in which BIRON did not react after the first utterance and the subject only repeated the verbal utterance without any movement.

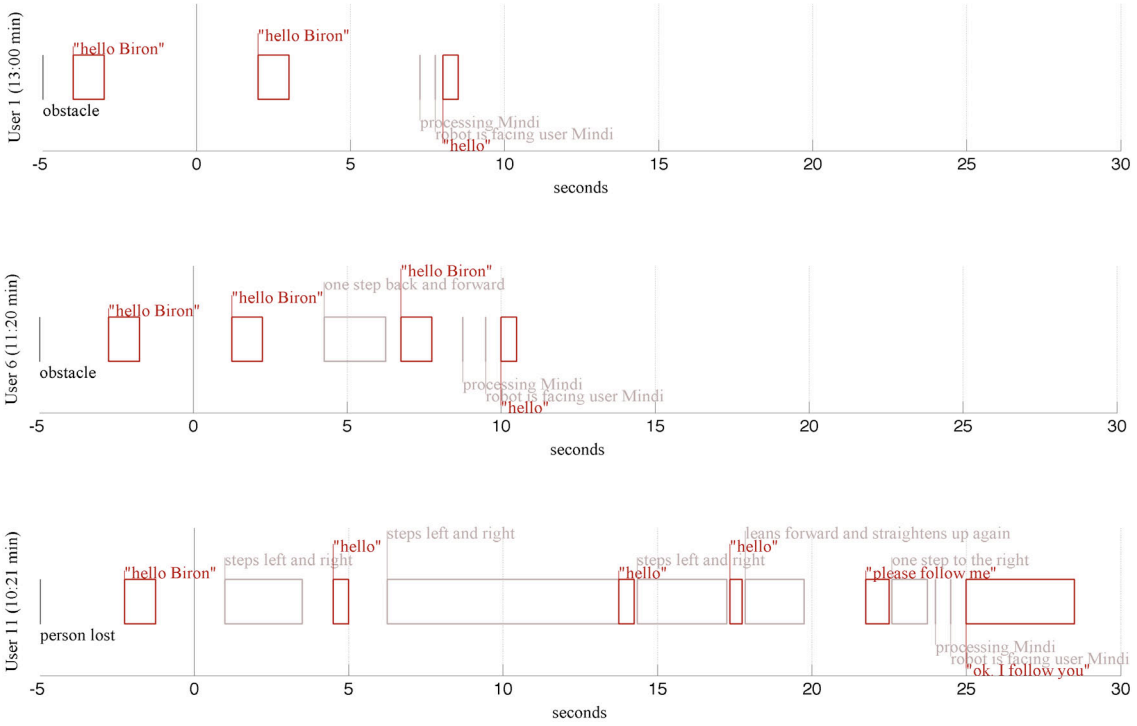


Figure 5-6. Strategies to attain the robot's attention with speech and gestures
The figure illustrates occurrences over time. In the upper row actions of the human are labeled, in the lower row actions of the robot. Red actions are verbal actions while light-grey actions are non-verbal actions (movements of the human, changes on BIRON’s display).

Strategy 2 contained five cases of five users all of which apart from this strategy used strategy 1 in other sequences. The rest of their sequences were short sequences as described above (saying “hello” or another command once and the interaction continued). In strategy 2, sequences contained saying “hello BIRON” or “BIRON hello”, movement, and at least one more verbal command. Interestingly, the movements of the users almost only consisted of stepping backwards and forwards in front of the robot (four instances). Only one person stepped to the left or the right (one instance). Moreover, in all sequences the exact same wording was repeated. The average length of sequences in this group was 18.25 seconds (minimum 8.75 seconds,

maximum 24 seconds measured from ‘poor person perception’ Mindi appearing to Mindi changing to ‘processing’).

Strategy 3 included six cases, three of which by one user (all her sequences) and one each of three other users. The strategy was characterized by the users moving after positioning themselves in front of the robot and before taking any verbal action. Altogether, the pauses between movements were shorter than between utterances. One user actually kept moving almost continuously with only four breaks in a sequence of 24 seconds. The average length of sequences with this strategy was 26.5 seconds (minimum 8.25 seconds, maximum 45 seconds) which is much longer than in strategy 1 and 2. In contrast to strategy 2, movement to the side is more common here (to the side - 20 times, backward/forward - 2 times, legs apart - 2 times, lean forward - 3 times). Only one person in one sequence stepped forward and backward. User 4 (twice in one sequence) and 11 (once) leaned forward and positioned their faces in front of the camera. They probably assumed that the robot had difficulties perceiving their faces but needed this percept in order to continue the interaction.

Strategy 1 was the most successful one in that the sequences were the shortest and the problem was resolved fastest. Verbal input by the users allowed the robot to identify them and to continue with the task. In accordance with expectation theory, the users repeated this successful strategy when the same situation reoccurred. Only two people gave the robot verbal feedback also if it failed at first. Their expectation seemed to be that BIRON needed a verbal percept which was probably furthered by many requests of the robot to say “hello” if something went wrong (for example, after moving the robot away from an obstacle, after having lost the percept of the person completely).

Users who applied strategy 2 also seemed to agree with the importance of greeting the robot to attract its attention even though the questionnaire data showed that their understanding of the Mindi picture differed. Only one user did not greet BIRON but started a new action (following). He was also the only one who moved during the utterance; all of the others only tried one action (speak, move) at a time. His movement was closely connected to the follow command. Therefore, with reference to the situational constraints, the function seems to be to start walking in the right direction rather than to attract attention. What is common to strategy 2 is that verbal utterances were repeated and enriched with movement. The direction of the movement in most cases was forward and backward. In contrast, the users who applied strategy 3 moved to the left and to the right. This movement was accompanied by a camera movement to both sides. Therefore, it was an obvious feedback that the robot was still doing something. This might have triggered subjects to try out more movements. No camera movement resulted when people walked backward and forward. This could be a reason that the participants who applied strategy 2 tried more verbal behavior.

Strategy 3 was usually only used once. Only one user repeated the strategy which was not successful with regard to the time needed to resolve the situation. However, it is noticeable that as the interaction proceeded she used a verbal utterance much earlier after the Mindi appeared. Obviously her expectations had changed at least in part. This finding supports the assumption that the history of the interaction influences expectations. In compliance with the model, the

users compare the behavior of the robot to their expectations. In the case described above, the robot disconfirmed the behavior of the user and she adapted accordingly.
Apart from this situation, another one shall be mentioned here shortly. If the participants could not get the robot to be attentive after a number of attempts, they used the reset command. A reset restarted the robot in a stable state and allowed the interaction to continue. However, also the usage of this command caused some concerns in the participants as the following examples show, which were retrieved with the off-talk analysis. One user asked if the robot could remember him after the reset and another one wanted to know whether the result of the reset would be that the robot drove back to the position where the interaction had started or if the robot would just restart the speech module. These utterances show that the effects of the reset were not quite clear. Hence, the participants did not have reliable behavior-outcome expectations. However, the command enabled the users to regain the robot's attention on the interaction level without a need for the experimenter to step in.

5.2.3 Ending the interaction with the robot (Systemic Interaction Analysis)

As was found for greeting, also parting follows a well-established script that people use many times a day. Laver (1975) has described parting sequences and their stages. Firstly the parting is initiated, for example, by an action like finishing a drink. Moreover, changes in proximity can occur. The initiator of the parting usually backs off a little to increase the distance between the participants. The other participant needs to accept this move which quite often does not happen. In this case the initiator usually resumes the degree of proximity he has moved away from and later again tries to initiate the closure. Laver (1975) stresses that the interaction can be closed only with mutual consent. Thus, the right feedback is necessary.
Next to changes in proximity, also changes in orientation and gaze direction are common in HHI. Before departing, the participants turn and gaze in the direction where they want to go to. They exchange verbal utterances that in the closing phase explicitly refer to the social and psychological aspects of the relationship between the participants (in contrast to greetings where neutral topics such as the weather might be discussed). Verbally, the participant who wishes to close the interaction gives a reason for parting (the reason might also concern the other participant "Don't you have to go to a class?") in order to consolidate the relationship between the two interactants. Such tokens can also carry implications of esteem for the other participant ("It was nice seeing you") or of caring for the other ("Hope your husband gets better soon", "Take care"). A second type of tokens refers to the continuation of the relationship ("See you later"). For example, in German this reference is also made with "Auf Wiedersehen". The formulaic phrases of farewell in English do not include this promise and fewer formulaic idioms are used. These kinds of tokens might also remind of social connections ("Say hello to Tom"). Additionally, conventional contact gestures of parting and conventional facial expressions might be exchanged. Thereafter, the distance between the participants begins to increase, a distant gesture of parting might be exchanged and, finally, the encounter is terminated by breaking the eye contact.

This description shows that there is a lot of room for variation depending on the situation in HHI. BIRON imitates part of this script reacting to the users' wish to close the interaction as specified in the *prototypical script*:

- user says bye (il) ("tschuess", "tschuess Biron", "auf Wiedersehen" ("bye", "bye Biron", "good bye")
- BIRON understands command correctly (sl)
- PTA changes from state person to state alertness (sl), person (ID) is logged out
- Mindi changes to thinking; thereafter to standing/empty (il)
- BIRON says ("Tschüss. Bis später" ["Bye. See you later"]) (il)
- BIRON sometimes adds "Die Interaktion hat mir viel Spaß gemacht" ["I liked the interaction very much"], "Ich verspreche, dass ich mich bessern werde" ["I promise that I will get better"])

The script shows that BIRON's behavior is restricted to the verbal utterances. The robot uses a formulaic idiom of farewell ("Tschuess" ["Bye"]) and a token that refers to the continuation of the relationship ("Bis später" ["See you later"]). Moreover, it can say something like "I liked the interaction very much" which signals esteem for the other participant, or "I promise I will get better" which is also a signal for continuation of the relationship.
The analysis of the data with SInA showed that 62.07% (18 out of 29) of the scenes followed this prototypical script. Three users did not finish the interaction with a farewell at all. All other deviations from the prototypical script are depicted in Table 5-14.
Table 5-14 shows one case which is quite typical for HRI. The robot always said "See you later". In some cases the users replied the same utterance which the robot could not understand. This disconfirmed the expectations of the users because in HHI people mostly use utterances that they also understand themselves. This example illustrates a severe problem in HRI. On the one hand, the robot's utterances need to be complex enough that the user knows what to do but, on the other hand, the robot's abilities to understand utterances are very restricted. However, the

Table 5-14. Deviation patterns in the farewell task

Pattern	User (il)	Robot (sl)	Robot (il)	Influence on HRI	#
Speech understanding					
Errors in speech recognition	utters farewell command	(a) the input cannot be interpreted at all	(a) asks user to repeat	(a) users repeat command	3
		(b) the input is interpreted in a wrong way	(b) starts a wrong action	(b) users try to resume farewell task	1
User speaks while robot speaks	utters farewell command while the robot speaks	partial speech recognition because the robot cannot listen while it talks	no reaction	users wait, after some time repeat the command	2
Repetition of commands	utters farewell command	processing/component is too slow	no reaction	users repeat command, robot answers twice	1
States					
Unfinished or wrong state	utters farewell command before completing some other action	does not accept farewell utterance; utterance is interpreted within the current system state	asks whether the user wanted to do sth. else, requests a stop command	users say stop to finish the previous task	4

results of the user study show that understanding the highly formalized reply “See you later” is necessary, especially because the robot uses it itself. Since it refers to the continuation of the relationship, it is useful to build up the social relationship which should be maintained. It is helpful especially because the robot does not signal departure with other modalities but speech. The users were not found to turn and gaze somewhere else before they parted as described by Laver (1975). However, in the prototypical case, the robot signals consent to the farewell verbally right away and, thus, the user does not need to initiate the closure in more complex ways. Moreover, the participants simply did not know where they had to go after the farewell.
The patterns person speaks while robot speaks and repetition of commands are similar to the patterns in the greeting sequence. The last deviation pattern, unfinished or wrong state, had not occurred during the greeting but was the most frequent pattern here. This problem arose when the robot was not in the person state when the user uttered the farewell because it had not finished some other action yet. Therefore, the robot first asked the users to quit the other action with a stop command before they repeated the farewell.
To conclude, it was found that the percentage of prototypical cases is similar to the greeting task which shows that both tasks follow well-established scripts that are similar in HHI. Thus, the users have clear behavior-outcome expectations. Also problems in the farewell sequence can easily be repaired. Again, this explains why the mean duration of the farewell sequences was rather short compared to other tasks (see Section 5.1.1).

5.2.4 Summary of the analyses of the social tasks of the home tour

The greeting and the farewell worked well (68%, 62%), especially compared to the functional tasks which the following section will show. Both social tasks follow a script that humans have practiced many times and the behavior of the robot seems to more or less comply to this script. Greeting was also found to be used to maintain attention. It was a part of all three strategies to secure the attention of the robot: verbal utterance only; verbal utterance enriched by movement; and movement enriched by verbal utterances.
Even though the scripts of greeting and farewell are rather straightforward, some deviation patterns were identified. Most deviations were connected to speech understanding (eight cases, 55%). In all but one of the cases the robot asked the user to repeat the utterance. This result is positive because asking for repetition solves the situation much more easily than starting a wrong task. Also that the user is not yet perceived before the interaction can rather easily be repaired with a greeting. This deviation was only evident at the beginning of the interaction and not during the farewell task. In contrast, the deviation pattern unfinished or wrong state only occurred in the farewell task.
Altogether, it can be noted that the participants at the beginning had some doubts about how to interact with the robot. However, they readily transferred their knowledge about the social tasks from HHI to HRI and performed them successfully. In other words, they developed target-based expectations about the robot based on their experience from HHI. This was possible because the users perceived the situations similarly both in HHI and HRI and used them to reach the same goals – to attain or maintain the attention of the opponent and to conclude the interaction.

5.3 Analysis of the functional tasks of the home tour

Three tasks were identified in Section 5.1.1 as being functional tasks: guiding the robot, teaching rooms, and teaching objects. The interaction was designed for the functional tasks and the users spent most time on them (see Section 5.1.1). In contrast to the social tasks they do not follow scripts in HHI that are as restricted as the scripts for greeting and parting. The following sections will show how this influenced the interaction and the deviations.

5.3.1 Guiding the robot (Systemic Interaction Analysis)

Guiding the robot is the task that enables the users to teach it different rooms and objects. Before being taught, the robot needs to be close to the object of interest. Therefore, the users initiate guidance tasks to place the robot and themselves in an appropriate location. Of course, the guiding task is not only a preparation for teaching but also a task of its own with difficulties, challenges, and its distinct prototypical script (see Lohse, Hanheide, Rohlfing, & Sagerer, 2009). The guiding task consists of three parts: giving the command to follow, guiding the robot, and saying stop to end the guiding sequence. As for all tasks, the subjects need to know what utterances the robot understands. Moreover, they have to estimate how far away they can be from the robot so that it still perceives them while following, but does not stop because the user is within the security distance (usually one meter for BIRON). The participants habitually expect that the spacing behavior of the robot is similar to human behavior and, thus, stand at a distance from the robot which is in accordance with an appropriate social space of European people (see, for example, Hüttenrauch et al. 2006; Walters et al., 2007). In contrast, some other concepts require training of the user. For example, saying stop at the end of a task is not something naturally done in HHI. Also the commands that the robot understands are not open ended. Therefore, the users have to learn during the interaction which utterances work well (here: “BIRON, folge mir” [“BIRON, follow me”] and “BIRON, komm mit” [“BIRON, come with me”]). On the system side, several prerequisites have to be fulfilled to conduct a follow behavior. The system needs to be in a consistent state in which all components accept the follow command and has to perceive the person stably. If this is the case, the *prototypical script* can proceed as follows:

- user says follow command (il) (“Biron folge mir” [“Biron follow me”], “Komm mit” [“Come with me”])
- BIRON understands the command correctly (sl)
- PTA changes to state follow, motor commands are enabled, and the respective components are started (sl),
- Mindi changes to follow (il)
- BIRON says “Gut, ich folge dir” [“Ok I follow you.”] [first trial]; “Ok. Ich folge dir. Bitte sage stopp wenn wir am Ziel sind” [“Ok I follow you. Please say stop when we have reached our destination.”] [second trial]) (il)
- user guides the robot (il)
- user says stop command (“Biron stop” [“Biron stop”], “Anhalten” [“stop”]) (il)
- BIRON understands commands correctly (sl)

- PTA changes to person (sl), the respective motor commands are sent and the respective components are stopped
- Mindi changes to standing/userlook (il)
- BIRON replies “Gut. Ich stoppe” (“Ok. I stop”)

The off-talk analysis showed that the users were very insecure about this script. They asked how far away from the robot they had to be for it to follow and when they should start moving. One user wanted to know whether it could follow in a certain direction and another one asked whether she was in the way of the robot when it started beeping. Moreover, they wondered if BIRON would stop right away when they did not move any more, when it encountered obstacles, or when asked to do so. All utterances showed insecurity about the robot’s abilities and the prototypical interaction script. The users were concerned especially with the robot’s ability to quickly stop in order not to run someone or something over. Of course, the robot should be able to stop quickly enough due to safety reasons.

Aside from security issues, it is not assumed that a prototypical situation is “perfect” or comparable to an ideal situation in HHI. Rather, it can be observed in the data that subjects explored many different strategies to successfully guide the robot. For example, they tried different positions in front of the robot to attract its attention and to keep an adequate distance that triggers the robot to follow. They stepped closer, further away, and to the side. They did not start guiding the robot by further increasing the distance until it had announced that it would follow or when it started driving, depending on which of these two actions happened first. Moreover, the data shows that the attention that the subjects paid to the robot varied, which is obvious in the orientation of the human body towards the robot. While some subjects always walked backwards to face the robot, others adapted their behavior according to the situation (see Section 5.1.3). When they were in open spaces such as the living room, these subjects turned their back to the robot and looked back at it once in a while. In contrast, when the path became narrow, they turned around and faced the robot the whole time. In the studies, it was also found to be prototypical that participants did not just walk around the corner assuming that the robot can interpret this like a human would do. Rather, they walked as far into the next room as possible without turning and waited until the robot came close before they turned.

In the next step, the cases in which the interaction deviates from this prototypical script will be analyzed for the 264 guiding sequences. 45 (17%) of these were categorized as prototypical interaction episodes. Consequently, 219 were deviating cases. This rate in fact is much higher than in the social tasks. With the help of the SInA method, twelve deviation patterns were identified which explain 98% (215) of the non-prototypical sequences. The deviation patterns can be categorized in four groups: speech understanding, person perception, state-related patterns, and navigation. In Table 5-15 they are analyzed according to what the users do, what happens within the system, what the robot does, the influence on the interaction, and the number of occurrences.

Four deviation patterns with respect to speech understanding were identified: speech recognition, repetition of commands, person speaks while the robot speaks, and speech fragments. Speech recognition errors are similar to the errors that occurred during the greeting

Table 5-15. Deviation patterns in the guiding task

Pattern	User (il)	Robot (sl)	Robot (il)	Influence on HRI	#
Speech Understanding					
Errors in speech recognition	utters follow and stop command	(a) the input cannot be interpreted at all	(a) asks the user to repeat	(a) users repeat the follow command	17
		(b) the input is interpreted in a wrong way	(b) starts a wrong action	(b) users try to resume guiding	14
User speaks while robot speaks	utters follow command while the robot speaks	partial speech recognition because the robot cannot listen while it talks	no reaction	users wait, after some time repeat the command	11
Repetition of commands	(a) utters follow command	processing/component is too slow	(a) no reaction	(a) users repeat follow command, robot answers twice, switches to follow state	7
	(b) utters stop command		(b) keeps driving until reaching the security distance, no verbal reaction	(b) users repeat command, robot answers twice, stops	9
Speech fragments	guides the robot without speaking	speech recognition catches fragments (a) the robot cannot interpret	(a) asks the user to repeat	(a) users repeat follow command or ignore robot utterance	10
		(b) the robot interprets in a wrong way	(b) starts a wrong action	(b) users try to resume guiding	4
Person Perception					
User lost (a) before following	utters a follow command, stands far away enough	no user perceived according to the person model	(a) does not react	(a) users wait, repeat command	22
(b) while following			(b) says that it cannot see the person anymore, requests a hello, Mindi displays weak person perception	(b) users say hello again, interaction continues	28
States					
Unfinished or wrong state	utters follow command before completing some other action	does not accept utterance; utterance is interpreted within the current system state	asks whether the user wanted to do sth. else, requests a stop command	users say stop to finish the previous task	27
Follow action incomplete	guides the robot, does not say stop, starts another action	stays in follow mode	asks whether the user wanted to do something else and requests a stop command	users say stop to finish the guiding task	13
Asynchronous dialog	says follow me, stands in an appropriate distance	has lost person, thinks that no interaction partner is present; the dialog reacts because it has a different expectation	verbally announces that it will follow, but does not start driving, Mindi displays follow state	users try to attract the robot's attention, walk around in front of the robot, try out another command after a while	8
Navigation					
Obstacle	guides the robot	an obstacle is perceived with the laser scanner	stops, announces obstacle and asks the user for help	users pull the robot away from the obstacle	34
User standing too close to the robot	guides the robot, stands within security distance	robot is in follow state, notices that person is standing within security distance	says that it follows, shows follow Mindi on screen, does not drive	interaction gets stuck if the users do not step back eventually	11

and the farewell tasks. Due to the errors, the speech could not be interpreted at all and the robot asked for clarification; or the speech was interpreted in a wrong way and the robot started an unexpected behavior (for example, the robot asked the users if they wanted to know where they were, instead of following). Clarification is rather natural in HHI, whereas starting wrong actions is not and takes more time to be repaired. Also repetition of commands and person speaks while robot speaks is similar to the patterns describe in Sections 5.2.1 and 5.2.3. A new pattern that occurred during guiding concerned speech fragments. Even though headsets were used for speech recognition, fragments occurred when someone talked in the surrounding of the robot or when the system interpreted some noise as speech. Speech fragments led to a request for repetitions or to wrong actions (for example, the system interpreted a "stop" and stopped following).

A functionality that is particularly relevant for guiding the robot is the robust tracking of the person. With SInA, two major deviation patterns were identified with this respect: person lost before following and person lost while following. In the first case, the robot did not react to the user because it hypothesized that nobody was there. In the second case, the robot stopped and announced that it could not see the user anymore. This was usually caused by subjects who walked too fast or too far to the side. This behavior again reflects the insecurity of the subjects about how to behave in this situation.

Like in the greeting and the farewell task, robot states also posed a problem in the guiding task. It considers unfinished or wrong states, incomplete follow actions, and asynchronous dialog. In contrast to unfinished or wrong state, which focuses on actions before the guiding sequence, the category incomplete follow action describes the incomplete follow action itself. Follow actions were not completed if the user did not say stop, but immediately tried the next action. Asynchronous dialog occurred when the dialog component had expectations that differed from the rest of the system. While the person perception did not perceive a person, this was not communicated to the dialog component that, as a result, reacted like a user was recognized and requested a verbal input. The input, however, was not considered because the system only interpreted verbal input if a user had been recognized. This particular improvement had already been made to the other tasks (see Peltason et al., 2009).

The last group includes two deviation patterns connected to navigation. These are obstacle blockage, and person standing too close. Obstacles, in general, are objects that the robot detects with the laser range finder within a certain security distance (approximately 30 cm measured from the center of the robot). They led to an emergency stop and to the announcement by the robot that an obstacle was in the way. Moreover, the user was asked for help ("Please pull me away from it"). If the users themselves were standing too close to the robot, it stopped when it was about to enter their personal space (about one meter for BIRON). In contrast to the obstacle blockage, in this case it did not announce anything but waited for the person to continue walking. If the person did not do so, the interaction got stuck.

5.3.2 Teaching rooms to the robot (Systemic Interaction Analysis)

The second functional task was to teach rooms to the robot. The SInA of this task is presented in Lohse, Hanheide, Pitsch, Rohlfing, and Sagerer (2009). Again, the prerequisites for this task are that the overall system state is consistent (all components can accept the command) and that the robot perceives the person stably. The *prototypical interaction* script of the task then looks as follows:

- user says room-teaching command (il) ("Biron das ist das <name of room>" ["BIRON, this is the <name of room>"])
- BIRON understands command correctly (both that a room is shown and the name of the room) (sl)
- PTA changes to state location-learning (sl)
- Mindi changes to looking at room (il)
- BIRON says "Das ist also das <name of room>" ("This is the <name of room> then") (il)
- BIRON conducts a 360° turn (il)

The robot conducts a 360° turn in order to acquire a decent representation of the room. This enables it to improve its metric mapping using SLAM (Guivant & Nebot, 2001). Furthermore, the robot uses its laser scanner to register the coarse layout of the room such as its size and its geometrical shape (Christensen & Topp, 2006).

As described before, it is not assumed that a prototypical situation is perfect or similar to an ideal situation in HHI. Rather, restrictions of the system are taken 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, it is tackled by providing appropriate feedback. Though the robot will usually not be able to track a person successfully, the idea is that feedback lets the users know how to re-initiate the interaction. Therefore, if the robot loses the users it asks them to say "hello" if they want to continue the interaction.

Altogether, 128 location-teaching 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 users and asked them to say "hello" if they wanted to continue the interaction (25). The remaining 92 cases (72%) were deviating cases. With the help of the SInA method, these were categorized into 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 all of which have been introduced with respect to other tasks: speech understanding, person perception, and state-related patterns. Table 5-16 depicts all patterns.

The only new pattern that occurred in the location-teaching task was *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

Table 5-16. Deviation patterns in the room-teaching task

Pattern	User (il)	Robot (sl)	Robot (il)	Influence on HRI	#
Speech understanding					
Errors in speech recognition	utters a room-teaching command	(a) input cannot be interpreted at all	(a) asks user to repeat	(a) users repeat command	5
		(b) input is interpreted in a wrong way	(b) starts a wrong action	(b) users try to resume task	31
Person perception					
User lost a) before room-teaching task	utters a room-teaching command	no user perceived according to person model	(a) does not react	(a) users wait, repeat command	1
b) during room-teaching task			(b) says that it cannot see the person anymore, requests a hello, Mindi displays weak person perception	(b) users say hello again; interaction continues	4
Third person	utters a room-teaching command	mistakenly classifies a third-person situation	does not react to the user	users wait; try to resume task	5
States					
Unfinished or wrong state	asks robot to learn a room before completing some other action	does not accept utterance; utterance is interpreted within the current system state	asks whether the user wanted to do sth. else, requests a stop command	users say stop to finish previous task	15
Action incomplete	teaches a room, does not wait for robot to turn before starting a new action	turning is interrupted	starts new action	robot cannot refine map, users do not notice deviation, no influence on HRI	25

only accepts verbal input from people facing it directly. This behavior is in line with the hypothesis that humans usually address others by looking at them. That is why BIRON did not react to commands if it detected a so-called "third person" that the other person might talk to. 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.

A very common pattern in the context of this task was action incomplete, i.e., the users did not allow the robot to turn in order to improve its representation of the room. This was due to the fact that the turning was not in line with the script of teaching rooms in HHI and the users did not learn in the interaction why the robot turned. Thus, they could not adapt their expectations to the situation. However, this deviation pattern did not have any impact on the interaction level and did not need to be repaired by the user.

5.3.3 Teaching objects to the robot (Systemic Interaction Analysis)

SALEM of the tasks has revealed that the object-teaching differed from the location-teaching with regard to many aspects (see Section 5.1). Moreover, the *prototypical interaction script* has to be differentiated for the trials. In the first iteration, the prototype was defined by the following actions:

- user says object-teaching utterance ("Biron das ist ein <name of object>" ["Biron, this is a <name of object>"]) and points out the object with a gesture (il)
- BIRON understands utterance (sl)
- Mindi switches to thinking/processing (il)
- PTA switches to state "object" (sl)
- Mindi switches to photo (il)
- BIRON looks at the object with its pan-tilt camera (il)
- BIRON says "das ist interessant, es gefällt mir gut" ("This is interesting, I like it") (il)
- Mindi changes to userlook (il)

However, the main problem with this prototype was that the robot feedback "That's interesting. I like it." was not credible. The users repeatedly asked whether the robot had really learned the object. Therefore, the prototypical interaction script was changed for the second trial as follows:

- user says object-teaching utterance ("Biron das ist ein <name of object>" ["Biron, this is a <name of object>"]) (il)
- BIRON understands utterance (sl)
- Mindi switches to thinking/processing (il)
- PTA change to state "object" (sl)
- BIRON answers "gut ich schaue mir das an" ("ok, I'm taking a look at it") (il)
- Mindi switches to photo (il)
- BIRON looks at the object with its pan-tilt camera (il)
- BIRON says "<name of object>, ich hab es mir angeschaut" ("<name of object>, I have seen it") (il)
- PTA change to state person (sl)
- Mindi changes to userlook (il)

Providing the user with the feedback of the object name made the interaction more credible. Unfortunately, in both iterations the gesture recognition of the robot only interpreted a few pointing gestures correctly. Therefore, the robot often looked in the wrong direction with its camera. This also explains why one user asked where the robot looked. However, since from the video recordings of the interaction the exact direction of the pan-tilt camera is hard to tell, also instances when the robot seemed to look in the wrong direction were treated as prototypical.
Altogether, 120 sequences were analyzed (55 from the first iteration, 65 from the second iteration). Only sequences in which the participants showed objects that they were asked to show by the experimenter were included. These were a chair and a table in first trials and a shelf and a floor lamp in the second trials. Some participants also showed other objects. These attempts often led to errors in speech understanding since the robot was not able to learn new words but only associated known words with objects in the room.
In the first iteration, 19 sequences (35%) were in line with the prototypical interaction sequence on the interaction level. However, in eleven of these, speech-understanding problems occurred, i.e., the robot understood a wrong object name. Because of the robot feedback in the respective

trial (“This is interesting. I really like it”), the participants did not notice these speech understanding errors. They would have noticed them in the second iteration, though. However, in the second trial no such error was evident. This might be due to the change of objects. The system better understood the words “Regal” (shelf) and “Lampe” (lamp) than “Stuhl” (chair) and “Tisch” (table). One reason for this might be that the words have two syllables instead of one. However, even though the error rate in speech understanding was reduced, the number of prototypical sequences was not much higher in the second trial (26 cases, 40%).

Table 5-17 shows the deviation patterns for both trials that classify 92% (69) of the deviations. All patterns that occurred have been explained in the context of the tasks that have been evaluated above.

As mentioned above, one additional problem in the object-teaching task was the gaze direction of the robot camera that was often inaccurate because it did not perceive the gestures correctly. In accordance with the theory, in both trials it was found that the credibility of the robot decreased if the robot did not look at the object (Argyle, 1969; Nagai, Asada, & Hosoda, 2006;

Table 5-17. Deviation patterns in the object-teaching task
(# 1/2 - number of occurrences in the first and in the second iteration)

Pattern	User (il)	Robot (sl)	Robot (il)	Influence on HRI	# 1/2
Speech understanding					
Errors in speech recognition	utters an object-teaching command	(a) input cannot be interpreted at all	(a) asks user to repeat	(a) users repeat command	3/13
		(b) input is interpreted in a wrong way	(b) starts a wrong action	(b) users try to resume task	18/5
User speaks while robot speaks	utters an object-teaching command while the robot speaks	partial speech recognition because the robot cannot listen while it talks	no reaction	users wait, after some time repeat the command	5/4
Repetition of commands	utters an object-teaching command	processing/ component is too slow	no reaction	users repeat command, robot answers twice	4/2
Person perception					
Third person	utters an object-teaching command	mistakenly classifies a third-person situation	does not react to the user	users wait, try to resume task	0/1
User lost (a) before object-teaching task	utters an object-teaching command	no user perceived according to the person model	(a) does not react	(a) users wait, repeat utterance	1/2
(b) during object-teaching task			(b) says that it cannot see the person anymore, requests a hello, Mindi displays weak person perception	(b) users say hello again, interaction continues	1/0
States					
Unfinished or wrong state	asks robot to learn an object before completing some other action	object cannot be learned, utterance is interpreted within the current system state	asks whether the user wanted to do sth. else, requests a stop command	users say stop to finish previous task	2/8

see Section 3.2.4). Often the users in this case showed the object again, even if the sequence was prototypical. Some also looked at the experimenter with a questioning facial expression.

5.3.4 Summary of the analyses of the functional tasks of the home tour

Altogether, the SInA of the functional tasks has revealed that these cause many more problems compared to the social tasks. While in the social tasks, more than 60% of the sequences were found to be prototypical, in the functional tasks many more deviations occurred. In the teaching tasks only 28% of the sequences for rooms and 38% of the sequences for objects were prototypical. The number for the guiding task was even lower with 17%. In all three tasks many deviations were caused by speech-understanding problems (guide: 33% of all deviations, room-teaching: 42%, object-teaching: 78%). In comparison to the guiding tasks, speech-understanding problems often caused wrong actions in the teaching tasks (rooms: 86%, objects: 43%, guide: 25% of all speech-understanding problems). This shows that guiding commands were not confounded with other commands. In contrast, in the teaching tasks, the robot often mistook rooms with objects and vice versa because both teaching commands followed the same sentence structure ("This is the …") and only the name of the object or room determines what was shown. Moreover, the robot often recognized utterances like "Biron, das ist das Wohnzimmer wo" ("Biron, this is the living room where"). The interrogative "where" led the robot to believe that the user wanted to know which room they were in. The probability of occurrence of this pattern depended very much on the speakers.

Looking at the percentages, speech understanding was least problematic in the guiding task (33% of all deviation patterns). All other types of deviation patterns were distributed equally for this task (person perception 21%, states 22%, navigation 23%). This result is positive with respect to person perception because it is difficult to track the users when they are moving and when the lighting conditions change. Nevertheless, the percentage is higher compared to the teaching tasks (room: 12%, object: 7%).

In the room-teaching task, most deviations were connected to states (47%). In 15 cases (17% of all deviation patterns) the state was unfinished or wrong and the user had to finish the previous action before teaching rooms. As the SALEM (see Section 5.1.1) has shown, teaching rooms is often preceded by the follow task. Since in the first trials the participants were not told that the guiding task had to be concluded with a "stop" command, this caused a problem. In the object-teaching task this deviation pattern included all deviations with respect to states (14% of all deviations). In the room-teaching task additionally the action itself was often incomplete which caused another 29% of all deviations. If looking again at the prototypical script of this task, it becomes obvious that the users did not understand the meaning of the 360° turn of the robot that it initiated to improve its representation of the room. Some people, therefore, interrupted the turn and the task was not finished adequately (see Lohse, Hanheide, Pitsch, Rohlfing, & Sagerer, 2009, for a concrete description of such a situation). However, this deviation is not severe because it did not have any consequences on the following actions and the users did not notice that the map was not refined.

To conclude, it can be noted that the robot has to communicate to the users why it performs certain behaviors or expects certain input, especially if the performed and the expected

behaviors are not in line with typical behaviors in HHI. The users need this information to establish target-based expectations that allow them to develop adequate behavior-outcome expectations. To reach this goal seems to be much harder for the functional tasks because they are more complex than social tasks and do not follow scripts that are well-established in HHI. Therefore, it is even more important that the robot's behavior supports the expectation formation of the users in a positive way.

5.4 Users' evaluation of the robot after the home tour interaction

This section reports the results of the questionnaires and the interviews that are a supplement to the interaction data. It serves to underline some of the insights gained above. The questionnaires and the interviews contained the same questions in both iterations of the study (see Appendix C). The comparison of the first and the second study is not the focus here. A first analysis showed that there were only slight differences between the ratings. These might be due to changes to the robot system (for example, the changed object-teaching prototype), however, they might also be attributed to differences between the users. This vagueness is one more reason not to speculate about the differences here.

A first positive result was that the users indicated that they liked BIRON (mean 4.01; all means on a scale of 1 [not at all] to 5 [very much]). However, they were not yet really satisfied with its behavior (mean 3.22). The following items describe reasons for this. First of all, the participants indicated that the robot was not very intelligent (mean 2.67) and 75% stated that it should be more intelligent. These numbers show that BIRON is far from seeming even close to being as intelligent as a human. The mean predictability of the robot was 3.34. However, with respect to this item the participants did not agree whether the robot should be more or less predictable. 54.17% argued it should be more predictable while the other 45.83% stated that it should be less predictable. Surprisingly, none of the participants was content with the degree of predictability. The answer to this question probably depended on how it was understood. The participants likely wish for less predictability on the interaction level, meaning that the robot uses varying utterances with the same meaning. In contrast, it can be assumed that all participants support a high degree of predictability with respect to the completion of the tasks. The same seems to be true for consistency (mean 3.59) where 31.82% argued that they wanted the robot to be more consistent and 68.18% favored less consistency. The only item that at least two participants (8.33%) thought was adequate in the study was talkativeness of the robot (mean 3.20). 33.33% indicated that the robot should be more talkative and 58.33% thought it should be less talkative. This difference cannot be attributed to different robot behavior in the trials. Therefore, it can be assumed that it was caused by the fact that the participants conceptualized the situation differently. If they wanted to solve the task as efficiently as possible, they wanted the robot to talk less as if they expect to be entertained during the interaction. In any case, most participants (87.50%) agreed that the robot should be faster than in the study (mean 1.95).

Even though the performance of the robot with respect to these items does not seem to be satisfactory to date, the participants indicated that the robot was very polite (mean 4.67), friendly (mean 4.63), attentive (mean 4.09), and cooperative (mean 4.00). This is probably one

reason that they liked it even though they found it a little boring (mean 2.63), not very useful (mean 3.14), and even less practical (mean 2.54).
In the last part of the questionnaire, the subjects had to rate statements about the interaction with BIRON. On the one hand, they indicated that they could hear well what the robot said (4.80). On the other hand, the robot had some problems understanding the users (mean 2.97) which led to a decreased fluency of the interaction (mean 2.50). On the positive side, the subjects did not think that the interaction with BIRON was extremely hard (mean 3.46) or frustrating (mean 2.45) and found it very easy to learn (mean 4.25). In general, they stated that BIRON was easy to handle (mean 3.76).
Next to the questionnaires, the participants also answered interview questions that were videotaped and annotated. The content analysis of the interviews revealed expectations of the users that underlined the findings of the questionnaires. The interviews mainly focused on the questions of what the users paid attention to in order to find out what the robot was doing, what they thought about the speech understanding and speech output, and what problem occurred in the interaction in general.
With respect to the question of what the users paid attention to, the answers showed that all of them concentrated on the screen, the speech output, the camera, and the actions of the robot (for example, driving). Most users interpreted the pan-tilt camera as the eye of the robot. However, they felt very differently about this as the following quotes show exemplarily:

> *"Am Anfang hatte ich Schwierigkeiten, da habe ich die Kamera als Auge aufgefasst und dann habe ich immer versucht, über das Auge Kontakt aufzunehmen, das war wahrscheinlich eine blöde Interaktion."*
> *("At the beginning I had problems because I understood the camera as being an eye and tried to make contact with the eye; this probably was a foolish interaction.")*

> *"Auch dieses Wahrnehmen fand ich gut, mit der Kamera. […] Das hat auch funktioniert, das heißt ich habe mich darauf eingestellt wo die Kamera steht und habe versucht, mich danach zu richten."*
> *("I also liked the perception with the camera. […] This also worked well, that is that I have adjusted to where the camera was and tried to act in accordance with it.")*

> *„Also ich finde es komisch, wenn er mich nicht anguckt"*
> *(„Well I find it weird when he does not look at me.")*

> *"wenn man in die Kamera guckt und man denkt sich, warum gucke ich jetzt in diese blöde Kamera? [...] Vielleicht sieht er Augenbewegungen oder so. Aber man hat irgendwie Angst davor, dass man ihn vermenschlicht und denkt, dass er reagiert, wenn ich gucke."*

(„When you look at the camera and think why am I looking at this stupid camera? [...] Maybe he sees the eye movement or something. But you're afraid that you'll humanize him and think that he reacts when you look.")

These examples show that the camera was interpreted as an eye. Some users really liked this while others felt strange about it because they did not want to anthropomorphize the robot. One user even thought that attracting the robot's attention via the camera was disadvantageous for the interaction. But in general these findings underline the questionnaire results that the participants found the robot very attentive. The thought that the camera did not really focus on the surrounding was probably related to the problems of the camera movement in the object-teaching task. Very often during this task the robot did not succeed in looking at the objects, which irritated the users as the following quote shows:

"Ich fände es gut, wenn er mehr auf Gestik reagiert oder auf eine Bewegung mit den Augen, dass man ihn besser so dirigiert. Weil er irgendwie sich dann immer auf das linke Regal fixiert hat, was ich ihm gar nicht gezeigt habe, weil ich das erst selber nicht gesehen habe. Ich habe ihm immer das rechte Regal gezeigt. Und darauf hat er nicht reagiert."
("I would appreciate it if he reacted more to gesture or eye movement as you can steer him better like that. Because he always concentrated on the left shelf that I hadn't shown to him because I hadn't seen it myself at first. I have always shown the right shelf. And he did not react to this.").

When this happened, the users often tried to maneuver the robot such that it could better perceive the object, which leads us to comments about problems that occurred in the context of guiding. The driving probably strongly contributed to the fact that the robot was perceived as being slow. One participant said in the interview:

"manchmal will man ihn einfach anpacken und dahin zerren wo man ihn braucht"
("Sometimes you just want to take the robot and drag it to where you need it.")

The participants commented that the navigation should be improved so that the robot can drive everywhere on its own and not need to be pulled away from obstacles. One person also mentioned that the robot should remember where he had last seen the person to be able to find her. Another participant stated that the user needed to be pretty far away from the robot before it started driving. Therefore, he thought, the robot could not take turns as it should.
With respect to speech recognition the judgments of the users again differed a lot. One person said that he was impressed that the robot understood spoken commands at all. In contrast, someone else commented that he had expected more of the speech understanding. Most of the other participants reported the concrete problems that they experienced with speech recognition and assumptions that they had developed about it. One participant thought that the robot did not

understand because she spoke too fast. She then tried to separate the words from each other. Also another user stated that she found the interaction exhausting because she had to speak loudly, slowly, and clearly.

Apart from the manner of speaking, some participants had also formed expectations regarding the content of what to say. One person said that he only used the word "stopp" ("stop") because if he said "Bitte anhalten" ("Please stop") the robot would run him straight over because it would not understand. Another person assumed that she was supposed to say stop after every misunderstanding and to start over again. Both assumptions are wrong; however, they do not influence the interaction in a negative way. They show that the users easily develop wrong expectations about the interaction. One user also communicated that she was very insecure about what to expect. She mentioned that when the robot had not understood an utterance, she had a hard time deciding whether she should repeat the same utterance or try to paraphrase it. She added that this restricts the verbal utterances one uses. Another participant underlined this by stating that she felt as if she talked to the robot like she would talk to an infant or a pet.

Even though the robot often misunderstood utterances, some participants seemed impressed about its ability to repair these situations:

> *"Ich hatte eigentlich erwartet, dass der Roboter einfach ausgeht, wenn ich was Falsches sage und stehenbleibt. Das war aber nicht der Fall."*
> *("Actually I had expected that the robot would turn off and stop when I said something wrong. This was not the case.")*

But even though the robot was able to repair the problems and eventually understand, it often failed to signal understanding in a way that the users believed that it had really learned something. This was especially true in the first session when the robot not yet repeated the name of the object that was shown. One user said that she was not sure whether the robot should have uttered something like "Oh, this is an armchair".

As in the questionnaires, the feedback about the robot's speech output in the interviews was more positive than the comments about the speech understanding. Most participants stated that they could understand the robot well. One said that the speech was "extremely monotonous" while others found it friendly. Another participant commented that the robot makes long pauses when speaking. Therefore, the users believed that it had finished and interrupted it. This problem could also be observed in the interaction data and is another reason that the robot was perceived as being very slow.

The participants had different opinions about what the robot said. While one person stated that she liked the robot's comments about the interaction (for example, "You are really good") and the rooms (for example, "You have a nice living room"), another one argued that the robot should not talk as much but rather concentrate on the tasks. This finding agrees with the questionnaire data.

To conclude, it can be noted that there is obviously no robot design that satisfies all users at the same time. It seems that the robots need to be adapted to single users and their expectations. In accordance with the model, this finding is due to the fact that each user perceives the physical

social situation individually, takes different contexts into account, and has different goals. As a result, the robot should be enabled to differentiate between the users, for example, it should find out whether the user is primarily focusing on task completion or enjoys the interaction as such. There are probably two approaches to this challenge: either the user defines the robot settings before the interaction or the robot quickly learns about its user's expectations and adapts itself.

5.5 Summary of the results of the SInA of the home tour studies

With the help of SInA, the five main tasks of the home tour have been analyzed and deviation patterns that occurred during the tasks were identified. The analysis has so far been separated for the social tasks and the functional tasks. It has been found that the social tasks are more clearly scripted and, therefore, caused fewer problems. The design of the tasks obviously confirmed the expectations of the users. In contrast, the functional tasks were more problematic. However, on the positive side, it has to be noted that all but three users completed the whole task on their own. In this section the deviation patterns that were identified in all tasks are summarized and compared to each other. Table 5-18 gives a first overview.

The table shows that speech recognition was the main problem in all tasks. Even though the users learned how to cope with this problem, they were not sure how to avoid it in the first place. As the questionnaires and the interviews have shown (see Section 5.4), the participants reasoned that they had to speak loudly and in a well-articulated manner, and often what they thought was a solution to the problem in fact worsened the situation. However, there was not really a way for the users to find out that their expectations were wrong because the robot did not tell them to speak differently. Moreover, this behavior seemed to be strongly anchored in the users because in HHI the same strategies are used if the opponent does not understand.

Also to interrupt others is something humans do to signal that they want to take the turn (Sacks, Schegloff, & Jefferson, 1974). This was also evident in the home tour studies and occurred in all tasks but teaching rooms. Unfortunately, the robot did not listen while it spoke and, thus, did not recognize the users' utterances. However, in the course of the interaction the users seemed to learn that this was the case and even though they kept talking while the robot talked, they repeated their utterances when it had finished.

Another behavior that the users could not suppress (in case they recognized that it led to a deviation pattern) was to repeat commands before the robot reacted. This deviation pattern was most evident in the guiding task. The users frequently repeated stop commands because they were not sure whether the robot had understood. This shows that the guiding task is more time-critical than other tasks. While the users readily wait for the robot to greet them, they want it to stop immediately in order not to run them or some object over.

Also speech fragments only occurred in the guiding task. They probably resulted from the noise caused by the movement of the human and the robot which the robot interpreted as some intended utterance it needed to react to. Thus, the robot asked the user to repeat the utterance, or in the worse case started a wrong action.

Table 5-18. Deviation patterns in all tasks
The numbers show how many times a deviation pattern occurred during the tasks.

Pattern	User (il)	Robot (sl)	Robot (il)	Influence on HRI	greet	farewell	guide	room	object
Speech Understanding									
Errors in speech recognition	utters a command	(a) input cannot be interpreted at all	(a) asks user to repeat	(a) users repeat command	4	3	17	5	16
		(b) input is interpreted in a wrong way	(b) starts a wrong action	(b) users try to resume task		1	14	31	23
User speaks while robot speaks	utters a command while the robot speaks	partial speech recognition because the robot cannot listen while it talks	no reaction	users wait, after some time repeat the command	1	2	11		9
Repetition of commands	utters a command	processing/component is too slow	no reaction	users repeat command, robot answers twice	1		16		6
Speech fragments	guides the robot without speaking	speech recognition catches fragments (a) the robot cannot interpret	(a) asks the user to repeat	(a) users repeat the follow command or ignore the robot's utterance			10		
		(b) the robot interprets in a wrong way	(b) starts a wrong action	(b) users try to resume task			4		
Person perception									
User lost (a) before action	utters a command	no user perceived according to person model	(a) does not react	(a) users wait, repeat command	5		22	1	3
(b) during action			(b) says that it cannot see the person anymore, requests a hello, Mindi displays weak person perception	(b) users say hello again, interaction continues			28	4	1
Third person	utters a command	mistakenly classifies a third-person situation	does not react to the user	users wait, try to resume task				5	1
States									
Unfinished or wrong state	utters command before completing some other action	does not accept utterance; utterance is interpreted within the current system state	asks whether the user wanted to do sth. else, requests a stop command	users say stop to finish the previous task		4	27	15	10
Action incomplete	starts another action without finishing the recent action	stays in recent mode	a) if stop is required, asks whether the user wanted to do something else and requests a stop command	a) users say stop to finish the current task			13		
			b) if no stop is required, starts the new action	b) no influence on HRI				25	
Asynchronous dialog	says follow me, stands in an appropriate distance	has lost person, thinks that no interaction partner is present; the dialog reacts because it has a different expectation	verbally announces that it will follow, but does not start driving, Mindi displays follow state	users try to attract the robot's attention, walk around in front of the robot, try out another command after a while			8		

Navigation					
Obstacle	guides the robot	an obstacle is perceived with the laser scanner	stops, announces obstacle and asks the user for help	users pull the robot away from the obstacle	34
User standing too close to the robot	guides the robot, stands within security distance	robot is in follow state, notices that person is standing within security distance	says that it follows, shows follow Mindi on screen, does not drive	interaction gets stuck if the users do not step back eventually	11
				sum	11 10 215 86 69

The next group of deviation patterns was connected to the state architecture. These deviation patterns are a sign that the users' expectations about the tasks did not match the actual task structure, i.e., they were not aware what actions the tasks consisted of. This is something that the robot needs to communicate if it differs from HHI (for example, demand a 'stop' command if this is not common in HHI).

Finally, navigation problems also only occurred during guiding. Certainly the robot should be enabled to follow on its own without being pulled by the users. The robot's abilities need to be improved with this respect, and already have been to a large degree.

5.6 Conclusion of the home tour studies

The analysis of the home tour studies has revealed many results that shall be summarized here. These findings concern the methodology as well as the interaction itself.

From the methodological point of view, SALEM was used for the analysis of gestures, gaze, and body orientation of the user towards the robot and the objects. For the analysis of gestures, a new coding scheme was promoted with respect to the teaching tasks. This was necessary because the gestures differed in the home tour studies compared to those conducted in the laboratory. While testing the coding scheme for interrater reliability, it was observed that the raters could not clearly differentiate whether the gestures were performed with the forearm or the whole arm. Therefore, it was concluded that the joints involved in the movement are of less importance for the description of the gestures than the intensity of the gesture which led to an adaptation of the coding scheme. Moreover, (un-) conventionalized gestures were coded in all other tasks.

Also with respect to body orientation, a coding scheme was developed. However, it turned out that this scheme needs further testing for reliability because with the current video data 45° shifts could often not be told apart accurately.

Next to SALEM, also SInA was applied here as a second main approach with the goal to analyze the task structure and to find reasons for the problems that arose in the interaction. Moreover, some data were visualized to show exactly how the users tried to maintain attention. The analysis was rounded off by data gained in questionnaires and interviews. Accordingly, this chapter has presented a broader amount of methods that shall be integrated here into one big picture that leads to the conclusion of the home tour studies.

In the object-teaching studies, positive and negative trials and the phases of the interaction were distinguished. In this chapter, tasks were compared to each other and each task was connected to different situations (either the task followed the prototypical script or it resulted in some

deviation pattern). The tasks that were identified belong to three groups: social, functional and problem-related tasks. The social tasks serve to direct the robot's attention. The functional tasks dominated the interaction. Thus, the home tour was found to be task-oriented, i.e., the users concentrated on teaching rooms and objects to the robot and to guide it around. The problem-related tasks interrupted these functional tasks. In the SInA they were not represented as tasks of their own but as deviation patterns. However, in the SALEM approach it was advantageous to treat them as tasks because this allowed to identify when certain problems occurred and to apply descriptive statistics.

Within the tasks, the modalities gaze, gesture, and body orientation were analyzed. During the teaching tasks the users more commonly gestured to teach objects rather than to teach rooms. The gestures that they used were mostly deictic which was also a result of the object-teaching studies. Thus, deictics were found to be most common in teaching tasks. In contrast, no iconic gestures were produced because either these were redundant in the situation or because the participants did not believe that the robot would understand them. Throughout the other tasks, it was evident that hardly any unconventionalized gestures were used and only some conventionalized gestures such as head nods and stop gestures. The usage of these particular gestures seemed highly habitual and automatic. Accordingly, in general it can be assumed that the users produced fewer gestures than in HHI because the robot did not gesture (McNeill, 1992). Thus, the robot's repertoire influenced the users' behavior. However, in the object-teaching studies, where the robot also did not gesture, all participants were found to gesture much more than the users in the home tour. This points to the fact that the task and the situation strongly influence the behavior. Apart from this, with respect to gestures in all studies it became most obvious that the behavior repertoires of the users strongly depend on personal factors.

In a next step, the body orientation of the users was analyzed. Again, it was clearly shown that the behaviors differed between the tasks which allows them to be distinguish from each other. The results revealed that the users spent most of the time with their upper body oriented toward the robot especially during the social tasks. The further they turned away, the shorter was the time that they spent in a certain orientation. Most switches in body orientation occurred in the functional tasks, especially during the guiding task. With respect to the teaching tasks, it was found that they differed because the users maintained a face-to-face orientation when teaching rooms but they turned to the side when teaching objects in order to set up an appropriate participation framework. However, these results have to be interpreted with caution because there are some uncertainties about the coding scheme and it needs to be verified whether changes of 45° can be reliably coded with a more adequate video recording. Furthermore, it could not be conclusively determined how the modality might structure the interaction. This question needs to be evaluated in a scenario that is more appropriate to distinguish communicative switches of body orientation from functional switches that are related to tasks such as guiding.

Also gaze was analyzed with SALEM. The results concerning gaze support the findings reported about body orientation. They underline that the users most often concentrated on the robot, especially in the social and the problem-related tasks. It was found that the glances at the robot were very long, longer than in HHI and also longer than in the object-teaching study. This

can be attributed to the fact that the users had to wait for a reply for a longer time because the robot was acting autonomously, and to the screen which was an additional modality that they looked at. Both facts again show that the robot behavior and the modalities influence the behavior of the users.

With respect to the tasks, SInA did not only reveal that they influenced the users' behavior repertoires but also that they caused different deviations. In general, it was shown that tasks with a well-established script in HHI that was adopted for HRI worked best. This is in line with the assumption that users transfer their expectations from HHI to HRI which helps them to develop appropriate behavior-outcome expectations. In other words, they know what to expect and how to behave in the situation.

The tasks with the best established scripts were the social tasks (greeting and farewell). During other tasks, the scripts were not as common in HHI and additionally differed in HRI. Most deviations were caused by the guiding task. One reason for this was that the robot did not sufficiently communicate the prototypical script of the task. If this is the case, the users easily develop wrong expectations that might lead to deviation patterns. Some of these were communicated by the users in the interviews (for example, attracting attention via the camera is disadvantageous; having to say stop after every task; speaking loudly and leaving pauses between all words of a sentence is helpful for the robot). With respect to the guiding task, one part of the script that was not clear to the users was that they needed to finish the task with a stop command. The task was not completed until they did so. In connection to this, the robot's repertoire should include strategies to credibly communicate task completion. This was also found for the object-teaching task where the credibility could be improved a lot if the robot at least repeated the name of the object after learning it. In general, if the robot communicated task completion, it supported the outcome predictions of the users and strengthened their expectations.

Many deviation patterns show that the users had wrong expectations or were not sure what to expect. Errors in speech recognition were caused by wrong expectations if the users said something that was outside of the robot's vocabulary. This rarely occurred in the home tour because most users kept using the utterances that they were trained to use. However, many of these deviations occurred because the users consciously adapted their manner of speaking and actually started to speak in a way that caused problems in the respective robot components. For example, they segmented the utterances as in "Biron <pause> This is a table". The word "Biron" alone led to a clarification question because the robot did not know how to handle it. Nevertheless, it needs to be mentioned that the robot was able to resolve the situations by uttering clarification requests. This is certainly an important component in the robot's behavior repertoire, especially because it does not have the same capabilities as a human and, thus, the interaction is asymmetric.

Another deviation pattern was caused by the users because they expected that they could reply while the robot was still speaking. The interviews have shown that often the participants felt that the robot had finished because there was a long pause between two parts of an utterance. Thus, two wrong expectations combined to the situation that the users tried to interrupt the robot

and it did not reply to what they said. Moreover, this situation was worsened by the users' feeling that the robot was too slow.

Another severe deviation pattern was that the robot lost the users. Mostly this was not caused by the user, but by environmental conditions. However, there are cases when the users walked too far away from the robot and the robot could not recognize them anymore. As the analysis of attention maintenance has shown, the users had different strategies to make sure that the robot is attentive. A positive result in this context was that all users recognized that the robot required a verbal input in order to recognize him or her.

Also the state-related deviation patterns unfinished/wrong state and incomplete action were caused by the users' inadequate expectations about the task structure. It has already been pointed out above that the robot has to communicate this structure or script.

Finally, the navigation-related deviations can in part be attributed to wrong expectations of the users. Especially the pattern user standing too close is definitely caused by inappropriate assumptions of the participants. In contrast, the obstacle problem was mostly caused by the robot's inability to pass through narrow doorways at the time of the study. However, the users' expectations with respect to guiding the robot also played a certain role, for example, if they expected that the robot could take a path that was as narrow as the one they took. Most of these constraints should be reduced by improving the navigation. Nevertheless, the robot also has to communicate what it can do, for example, what are the prerequisites for it starting to drive. This communication can be achieved with speech but also with screen output (Mindi) or other modalities. Moreover, the example of the obstacle showed that the robot's abilities need to be in line with its role. In the interview, it became clear that the participants were ready to help the robot by pulling it; however, they commented that an actual assistant in the household should not depend on the assistance of the user.

All the disadvantageous expectations reported so far cause problems in the interaction. However, not all wrong expectations necessarily cause problems; for example, if the users said stop after every task, this slowed down the interaction, but apart from that the interaction continued as intended.

To conclude, the analysis revealed insights about the behaviors of the users and about the expectations that are connected to them. In accordance with the model, these expectations strongly depend on the situation, the task, and the robot behavior. SInA has shown which deviations from a prototypical script were caused by user expectations that are not in line with the design of the robot BIRON that was evaluated. In a next step, these insights allow to adapt the robot to the users' expectations, or to enable it to communicate its own expectations in cases where an adaptation is not possible.

6 Conclusion

This thesis started out with four main goals (see Section 1.3). The first goal was to develop a model that describes the dynamic changes of expectations in the interaction situation and supports the prediction of the user behavior. This goal was accompanied by the aim to develop an approach to provide an in-depth description of the situation as such. The third aim concerned the development of qualitative and quantitative methods to systematically research questions that come up in the context of HRI. The fourth goal was to show that the findings contribute to research on expectations in general and not only in HRI. This section summarizes how these goals were reached and what questions need to be addressed in future research.

Expectation- and situation-based model for HRI

The first aim of the thesis was to develop a model that describes HRI as an interaction situation that influences the expectations of the user. This model has been introduced in Section 2.3. It is based on the theoretical background of literature about the concepts of expectation and situation. The expectations are described in the model as a result of the perception of the situation by the user who is influenced by the context and the goals. The goal of the model was to capture the dynamic changes of the expectations of the users in order to be able to better predict their behavior. Indeed, various analyses have shown that the robot confirmed or disconfirmed the users' expectations which led to certain behaviors of the human. Since many participants performed the same behaviors, it could be concluded that they were caused by the situation. Thus, it has been shown that the model in fact allows to predict the users' behavior in certain situations.

Description of the HRI situation

Since the situations are one main factor of this model, an approach was needed to describe them in depth in order to compare them to each other and to determine how they change and how the changes influence the expectations and the behavior of the users. This was the second aim of the thesis.

The description of the situation should include a thorough analysis of the behavior of the users. With respect to this aim, features from HHI (Argyle, Furnham, & Graham, 1981) were transferred to HRI to characterize the situation (see Section 2.1.1.3). Moreover, an approach was introduced to analyze the situation on different levels of specificity (see Section 2.1.1.3). It has been shown that the analysis profits from taking the multimodality of the interaction into account.

In the data analysis sections the approach was used to describe interaction situations on the scenario level, on the task level, and on the level of specific situations (highest degree of specificity). Also, HRI situations could be compared to HHI situations on a lower level of specificity. Thus, an approach was found to differentiate interaction sequences on different levels of complexity and to show that the users' behavior changes during the interaction when the situation changes. Every situation requires the usage of certain elements of the behavior repertoires in a certain sequence. If these expectation-based behaviors and the sequences are

known to the robot, it might in fact be enabled to better understand what the users are trying to do and its situation awareness might improve. In other words, the robot might learn to better predict the users' behavior based on its knowledge about past situations.

Methods that contribute to HRI research

The third aim of this thesis concerned the development of qualitative and quantitative methods to systematically research HRI. In Chapter 3, methods were introduced to reach this goal. The importance of coding schemes to identify behaviors of the users was stressed. Coding schemes were developed for gesture, speech, and body orientation. SALEM was introduced to analyze the annotations of the behaviors that are based on the coding schemes. It has been shown to fulfill the criteria of generalizability and applicability to all kinds of modalities. Most importantly, it allows for efficient quantitative analysis. This was shown by the variety of SALEM results that could be presented in this thesis. The approach put all these results on the same statistical grounds which increased consistency.

Another method that has been developed in the course of this thesis is SInA. Its role is to analyze the interaction on the task level and to identify deviations from prototypical interaction scripts. SInA has been shown here to be applicable to a variety of tasks. It can be used to compare them to each other with respect to the kind of deviations and amount of deviations that they result in. These deviations have been revealed to be based on disadvantageous behavior of the users caused by wrong expectations or by problems on the system level. The deviations that were identified on the system level support the concrete improvement of the system components.

These key methods were enriched with visualization of data in order to identify the exact timing and sequences of the users' behaviors. Moreover, the participants' questions about the interaction, questionnaire ratings, and impressions about the interaction were presented to underline the inferences that were drawn from the other analyses.

In summary, the thesis has introduced a variety of methods that allow HRI to be analyzed from different points of view. The usage of the methods has been shown here with respect to the first aim of the thesis, to analyze the behavior of the users in different situations and to infer the expectations that led to this behavior.

Contribution to the literature about expectations

The fourth aim was to advance the literature about expectations by way of insights from HRI and to show that HRI is a valuable research field for the expectations research. In order to achieve this aim, expectation theory was introduced to HRI (see Section 2.2). The thesis has focused on probabilistic expectations. Many examples were given that showed specific expectations of the users in HRI. In various places, it could be shown that these expectations were based on knowledge from HHI. However, the robot often violated the expectations because it had fewer abilities and less knowledge than the human and the interaction was asymmetric. Decreasing this asymmetry is directly related to the first aim to develop a model to better predict the users' behavior and to increase the robot's knowledge about the situation. However, it can also be exploited to deliberately research the evolvement of the expectations in

asymmetric interaction. In this context, the robot can be used to produce the same behavior over and over again in a similar way. Therefore, the changes in user behaviors and expectations can be directly attributed to a deterministic behavior of the robot which is a great advantage of HRI. It has been demonstrated that the influences of certain robot behaviors can be analyzed with respect to the formation of target-based expectations and the reaction to expectation violations. Thus, HRI has been shown here to contribute to the literature about expectations, as well as it has been proven that the expectation theory from the social sciences advances HRI.

Discussion and future perspectives

This thesis has shone light on the question of the influence of expectations on HRI. It has been demonstrated that expectations influence the behavior of the users. Knowing these expectations and enabling the robot to perceive them will improve the robot's abilities to understand the situation. On the other hand, also the robot has expectations regarding input of the user and regarding environmental requirements. These expectations are based on technical details of the implementation. Often the users do not know them because they differ from expectations in human interaction. That is why another viewpoint on the topic that needs to be taken in future research is the robot's expectations. The main question is how the robot can communicate its own expectations in an understandable way without requiring the user to have a lot of knowledge about the system. Understanding the robot's expectations should be easy for the users in order to enable them to develop correct expectations about the robot and to improve their own behavior-outcome expectations. At the moment, often the users have wrong expectations about the systems. However, it has also been shown that they adapted their expectations depending on the situation. The adaptations are based on their own reasoning about the situation. Certainly this reasoning might change in long-term interaction and in other situations. Therefore, another main aim of future research should be to evaluate the concept of expectations and the model developed in this thesis in the context of longer-term studies and different situations. One aspect that could be of importance is the recognition of emotional aspects of the users' expectations. This is useful to improve the robot's ability to recognize when the expectations of the users were (dis-) confirmed and to react appropriately. Also the comparison of tasks embedded in different scenarios should be in the interest of researchers in order to find general guidelines for robot design.

Further analyses could easily build on the methods introduced in this thesis which are generalizable and applicable to different scenarios, situations, and research questions. They have served here to analyze the behavior of the users and the tasks that needed to be solved. However, the SALEM and the SInA approach could also be used to conduct more analyses on the system level. Hence, the analyses could serve even more to improve the components of the robot. In fact, this should probably be the most important goal of future work. The tools and analyses introduced in this thesis should be used to realize changes that positively affect the robot's components and behaviors in order to actually improve the interaction. This is a strongly interdisciplinary task that constitutes the core challenge of HRI today and in the future.

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.*

Aiello, J. (1987). Human spatial behavior. In D. Stokols & I. Altman (Eds.). *Handbook of Environmental Psychology 1*. New York: Wiley Interscience, 505-535.

Ajzen, I. (1988). *Attitudes, personality, and behavior*. Chicago: Dorsey Press.

Allwood, J. (2001). Dialog Coding - Function and Grammar: Göteborg Coding Schemas. *In Gothenburg Papers in Theoretical Linguistics 85*, University of Göteborg, Dept of Linguistics, 1-67.

Anderson, C. A. (1983). Imagination and Expectation: The Effect of Imagining Behavioral Scripts on Personal Intentions. *Journal of Personality and Social Psychology, 45(2)*, 293-305.

Annett, J. & Stanton, N. A. (2000). Research and developments in task analysis. In J. Annett & N. A. Stanton (Eds.). *Task Analysis*. London and New York: Taylor & Francis, 1-8.

Argyle, M. (1969). *Social Interaction*. London: Methuen & Co Ltd.

Argyle, M. (1975). *Bodily communication*. New York: International Universities Press.

Argyle, M. (1988). *Bodily Communication*. London: Methuen.

Argyle, M., Furnham, A., & Graham, J. (1981). *Social Situations*. Cambridge: Cambridge University Press.

Arkin, R. M. & Duval, S. (1975). Focus of Attention and Causal Attribution of Actors and Observers. *Journal of Experimental Social Psychology, 11*, 427-438.

Aronson, E., Wilson, T. D., & Akert, R. M. (2003). *Social Psychology*. Upper Saddle River, NJ: Pearson Education.

Arras, K. O. & Cerqui, D. (2005). *Do we want to share our lives and bodies with robots? A 2000-people survey.* Technical Report Nr. 0605-001. Autonomous Systems Lab Swiss Federal Institute of Technology, EPFL.

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2006). The Influence of People's Culture and Prior Experiences with Aibo on their Attitude Towards Robots. *AI & Society. The Journal of Human-Centred Systems, 21.* 217-230.

Batliner, A., Hacker, C., & Nöth, E. (2006). To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk. In *How People Talk to Computers, Robots, and Other Artificial Interaction Partners,* SFB/TR 8 Report No. 010-09/2006*,* 79-100.

Bierbrauer, G. (2005). *Sozialpsychologie*. Stuttgart: Kohlhammer.

Biernat, M. (2005). *Standards and expectancies: contrast and assimilation in judgments of self and others*. Hove: Psychology Press.

Blanck, P. D. (Ed.) (1993). *Interpersonal Expectations. Theory, research, and applications.* Cambridge: Cambridge University Press.

Bohner, G. & Wähnke, M. (2002). *Attitudes and attitude change*. Hove: Psychology Press.

Brand, R. J., Shallcross, W. L., Sabatos M. G., & Massie, K. P. (2007). Fine-Grained Analysis of Motionese: Eye Gaze, Object Exchanges, and Action Units in Infant-Versus Adult-Directed Action. *Infancy, 11(2)*, 203-214.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems 42,* 167-175.

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

Burgoon, J. (1983). Nonverbal Violations of Expectations. In Wiemann, J. & Harrison, R. (Eds.). *Nonverbal Interaction*. Sage Annual Reviews of Communication Research, 77-111.

Burgoon, J. K. (1993). Interpersonal Expectations, Expectancy Violations, and Emotional Communication. *Journal of Language and Social Psychology,12*, 30-48.

Burgoon, J. K. & Jones, S. B. (1980). Towards a Theory of Personal Space Expectations and Their Violations. In B. W. Morse & L. A. Phelps (Eds.). Interpersonal Communication: A Relational Perspective *Human Communication Research,* 198-212. [reprint, original in *Human Communication Research, 2(2)* (1976)].

Butterworth, G. (1992). Context and cognition in models of cognitive growth. In P. Light & G. Butterworth (Eds.). *Context and Cognition: Ways of Learning and Knowing*. London: Prentice Hall, Harvester Wheatsheaf, 1-13.

Butterworth, B. & Beattie, G. (1978). Gestures and silence as indicator of planning in speech R. Campbell & P. Smith (Eds.). *Recent Advances in the Psychology of Language: Formal and Experimental Approaches*. New York: Olenum Press, 347-360.

Cassell, J. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill, (Eds.). *Embodied Conversational Agents*. Cambridge, MA: MIT Press*,* 1-27.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). Non-Verbal Cues for Discourse Structure. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics,* 114-123.

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

Clark, H. H. (2003). Pointing and Placing. In S. Kita (Ed.). *Pointing. Where language, culture, and cognition meet*. Mahwah, NJ/London: Lawrence Erlbaum Associates, *243-268*.

Cole, M. & Cole, S. (1989). *The development of children*. New York: Scientific American*.*

Craik, K. H. (1981). Environmental Assessment and Situational Analysis. In D. Magnusson (Ed.). *Towards a Psychology of Situations.* Hillsdale, NJ: Lawrence Erlbaum Associates*,* 37-48.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working Minds. A practitioner's guide to cognitive task analysis.* Cambridge, MA: MIT Press.

Darley, J. & Batson, C.D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behaviour. *Journal of Personality and Social Psychology, 27*, 100-108.

Darley, J. M. & Fazio, R. H. (1980). Expectancy Confirmation Process Arising in the Social interaction Sequence. *American Psychologist, 35*, 867-881.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philosophical Transactions of The Royal Society B, 362*, 679-704.

Dautenhahn, K., Ogden, B., & Quick, T. (2002). From embodied to socially embedded agents - implications for interaction-aware robots. *Cognitive Systems Research. (Special Issue on Situated and Embodied Cognition), 3(3)*, 397-428.

Dautenhahn, K., Walters, M., Woods, S., Nehaniv, K. K. C., Sisbot, A., Alami, R., & Siméon, T. (2006). How may I serve you?: a robot companion approaching a seated person in a helping context. *Proceedings of the 1st ACM SIGCHI/SIGART conference on human-robot interaction,* 172-179.

Dautenhahn, K. & Werry, I. (2002). A Quantitative Technique for Analysing Robot-Human Interactions. *Proceedings of the 2002 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems; EPFL, Lausanne, Switzerland,* 1132-1138.

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M., Koay, K., & Werry, I. (2005). What is a robot companion - Friend, assistant or butler?. *Proc. IEEE IROS, (Edmonton, Canada, 2005),* 1488-1493.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (2004). *Human-Computer Interaction.* Harlow: Pearson Education Limited.

Eagly, A. H. & Chaiken, S. (1993). *The psychology of attitudes.* Orlando, FL: Harcourt Brace Jovanovich College Publishers.

Endsley, M. R., Bolté, B., & Jones, D. G. (2003). *Designing for Situation Awareness. An Approach to User-Centered Design.* London and New York: Taylor & Francis.

Endsley, M. R. (1988). Design and Evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society 32nd Annual Meeting.* Santa Monica, CA: Human Factors Society, 97-101.

Engelhardt, R. & Edwards, K. (1992). Human-robot integration for service robotics. In M. Rahimi & W. Karkowski (Eds.). *Human-Robot Interaction.* London: Taylor & Francis*,* 315-347.

Fasel, I., Deak, G., Triesch, J., & Movellan, J. (2002). Combining embodied models and empirical research for understanding the development of shared attention. *Proceedings of the 2nd International Conference on Development and Learning*, 21-27.

Fazio, R. & Zanna, M. (1981). Direct experience and attitude-behavior consistency. In L. Berkowitz (Ed.). *Advances in experimental social psychology, 14.* San Diego, CA: Academic Press, 161-202.

Feather, N. T. (1982a). Action in Relation to Expected Consequences: An Overview of a Research Program. In N. T. Feather (Ed.). *Expectations and Actions: expectancy-value models in psychology*. Hillsdale: Lawrence Earlbaum Associates, 53-95.

Feather, N. T. (1982b). Expectancy-Value Approaches: Present Status and Future Directions. In Feather, N. T. (Ed.). *Expectations and Actions: expectancy-value models in psychology*. Hillsdale: Lawrence Earlbaum Associates, 395-420.

Feather, N. T. (1982c). Introduction and Overview. In N. T. Feather (Ed.). *Expectations and Actions: expectancy-value models in psychology*. Hillsdale: Lawrence Earlbaum Associates, 1-14.

Fischer, K. (2000). *What is a situation?* Gothenburg Papers in Computational Linguistics, *00-05,85-92.*

Fischer, K. & Lohse, M. (2007). Shaping Naive Users' Models of Robots' Situation Awareness. *Proceedings of IEEE International Conference on Robot & Human Interactive Communication (RO-MAN'07).* Jeju, Korea.

Fishbein, M. & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Reading, MA: Addison-Wesley.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42 (3-4)*, 143-166.

Fong, T., Thorpe, C., & Baur, C. (2001). Collaboration, Dialogue, and Human-Robot Interaction. *In International Symposium on Robotics Research (ISRR).* Advanced Robotics Series, Springer.

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching Robot Appearance and Behavior to Tasks to Improve Human-Robot Cooperation. *Proceedings of IEEE International Conference on Robot & Human Interactive Communication (RO-MAN'03),* 55-60.

Goffman, E. (1961). *Encounters*. Indianapolis: Bobbs-Merrill.

Goldin-Meadow, S. (2003). *Hearing gesture. How our hands help us think.* Cambridge, MA: The Belknap Press of Harvard University Press.

Goodrich, M. A. & Schultz, A. C. (2007). Human-Robot Interaction: A Survey. *Foundations and Trends in Human-Computer Interaction, 1(3)*, 203-275.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York, USA: Academic Press.

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

Goodwin, C. (2003). Pointing as Situated Practice. In S. Kita (Ed.). *Pointing: Where Language, Culture, and Cognition Meet*. Mahwah, NJ; London: Lawrence Erlbaum Associates Publishers*,* 217-241.

Green, A. (2009). *Designing and Evaluating Human-Robot Communication. Informing Design through Analysis of User Interaction.* Doctoral Thesis. KTH Computer Science and Communication, Stockholm, Sweden.

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.

Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge, MA: Cambridge University Press.

Hackos, J. T. & Redish, J. C. (1998). *User and task analysis for interface design*. New York: Wiley and Sons.

Hall, E. (1963). A System for the Notation of Proxemic Behavior. *American Anthropologist, 65*, 1003-1026.

Hall, E. (1966). *The Hidden Dimension: Man's Use of Space in Public and Private*. London, UK: The Bodley Head Ltd.

Hanheide, M. & Sagerer, G. (2008). Active Memory-based Interaction Strategies for Learning-enabling Behaviors. *Proc. International Symposium on Robot and Human Interactive Communication (RO-MAN'08).*

Harris, M. & Rosenthal, R. (1985). Mediation of Interpersonal Expectancy Effects: 31 Meta-Analyses. *Psychological Bulletin, 97 (3)*, 363-386.

Hatano, G. & Inagaki, K. (1992). Desituating cognition through the construction of conceptual knowledge. In P. Light & G. Butterworth (Eds.) *Context and Cognition*. London: Prentice Hall, Harvester Wheatsheaf. 115-133.

Hayduk, L. (1983). Personal space: Where we now stand. *Psychological Bulletin, 94*, 293-335.

Heckhausen, H. (1977). Achievement motivation and its constructs: A cognitive model. *Motivation and Emotion, 1 (4)*, 283-329.

Heckhausen, J. & Heckhausen, H. (2006). Motivation und Handeln [Motivation and Action]. Heidelberg: Springer Medizin Verlag.

Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008). Theory of Mind (ToM) on Robots: A Neuroimaging Study. *The Third ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008),* Amsterdam, the Netherlands.

Hegel, F., Lohse, M., Swadzba, A., Wachsmuth, S., Rohlfing, K., & Wrede, B. (2007). Classes of Applications for Social Robots: A User Study. *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN'07),* Jeju Island, Korea.

Hegel, F., Lohse, M., & Wrede, B. (2009). Effects of Visual Appearance on the Attribution of Applications in Social Robotics. *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN'09),* Toyama, Japan.

Hirt, E. T. (1990). Do I See Only What I Expect? Evidence for an Expectancy-Guided Retrieval Model. *Journal of Personality and Social Psychology, 58 (6)*, 937-951.

Hirt, E., Lynn, S., Payne, D., Krackow, E., & McCrea, S. (1999). Expectancies and Memory: Inferring the past from what must have been. In I. Kirsch (Ed.). *How expectancies shape experience*. American Psychological Association, 93-124.

Hüttenrauch, H., Severinson-Eklundh, K., Green, A., & Topp, E.A. (2006). Investigating Spatial Relationships in Human-Robot Interaction. *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems,* Beijing, China.

IFR statistical department (2007). *The Robots are coming!.* Press release. October, 23, 2007.

Iverson, P., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother–child interaction. *Cognitive Development, 14*, 57-75.

Jones, E. E. (1990). *Interpersonal Perception*. New York: W.H. Freeman and Company.

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

Kaplan, F. (2005). Everyday robotics: robots as everyday objects. *sOc-EUSAI '05: Proceedings of the 2005 joint conference on Smart objects and ambient intelligence, ACM,* 59-64.

Kelley, H. H. (1950). The warm-cold variable in first impressions of persons. *Journal of Personality, 18*, 431-439.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

Kendon, A. (1990). *Conducting Interaction – Patterns of Behavior in Focused Encounters.* Cambridge, MA: Cambridge University Press.

Kendon, A. (2004). *Gesture. Visible Action as Utterance.* Cambridge, MA: Cambridge University Press.

Khan, Z. (1998). *Attitudes towards intelligent service robots.* Nr. TRITA-NA_P9821, IP-Lab-154. IPLab, Royal Institute of Technology (KTH), Sweden.

Kiesler, S. B. (1973). Preference for predictability or unpredictability as a mediator of reactions to norm violations. *Journal of Personality and Social Psychology, 27*, 354-359.

Kim, H. & Kwon, D. (2004). Task Modeling for Intelligent Service Robot using Hierarchical Task Analysis. *Proc. of the 2004 FIRA Robot World Congress,* Busan, Korea.

Kirwan, B. & Ainsworth, L. (1992). *A Guide to Task Analysis*. London: Taylor and Francis.

Koay, K. L., Syrdal, D. S., Walters, M. L., & Dautenhahn, K. (2007). Living with Robots: Investigating the Habituation Effect in Participants' Preferences During a Longitudinal Human-Robot Interaction Study. *Proceedings IEEE International Conference on Robot & Human Interactive Communication (ROMAN'07).* Jeju, Korea, 564-569.

Koay, K., Walters, M., & Dautenhahn, K. (2005). Methodological Issues Using a Comfort Level Device in Human-Robot Interaction *Proceedings IEEE International Conference on Robot & Human Interactive Communication (ROMAN'05).*

Kopp, S. (2006). *How People Talk to a Virtual Human - Conversations from a Real-World Application*. How People Talk to Computers, Robots, and Other Artificial Interaction Partners, SFB/TR 8 Report No. 010-09/2006*,* 101-111.

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.

Lang, C., Hanheide, M., Lohse, M., Wersing, H., & Sagerer, G. (2009). Feedback Interpretation based on Facial Expressions in Human–Robot Interaction. *Proceedings of 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'09).*

Lang, S., Kleinehagenbrock, M., Hohenner, S., Fritsch, J., Fink, G. A., & Sagerer, G. (2003). Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. *Proc. Int. Conf. on Multimodal Interfaces*, Vancouver, Canada, 28-35.

Langer, E. J. & Abelson, R. F. (1974). A patient by any other name: Clinician group difference in labeling bias. *Journal of Consulting and Clinical Psychology, 42*, 4-9.

Laver, J. (1975). Communicative Functions of Phatic Communion. In A. Kendon, R. M. Harris, & M. R. Key (Eds.). *Organization of Behavior in Face-to-Face Interaction*. The Hague: Mouton & Co*,* 215-240.

Lee, M. K. & Makatchev, M. (2009). How Do People Talk with a Robot? An Analysis of Human-Robot Dialogues in the Real World. *Proceedings CHI 2009 ~ Spotlight on Works in Progress ~ Session 1,* 3769-3774.

LePan, D. (1989). *The cognitive revolution in western culture: The birth of expectation*. London: MacMillan Press.

Lewin, K. (1935). *A Dynamic Theory of Personality: Selected Papers*. New York, London: McGraw-Hill.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves. *Pers Soc Psychol Bull, 30(9)*, 1175-1185.

Lohse, M. (2009). Expectations in HRI. *Proceedings Workshop New Frontiers in Human-Robot Interaction, AISB'09,* Edinburgh, UK.

Lohse, M., Hanheide, M., Pitsch, K., Rohlfing, K.J. & Sagerer, G. (2009). Improving HRI design applying Systemic Interaction Analysis (SInA). *Interaction Studies 10(3),* 299-324.

Lohse, M., Hanheide, M., Rohlfing, K., & Sagerer, G. (2009). Systemic Interaction Analysis (SInA) in HRI. *The Fourth ACM/IEEE International Conference on Human-Robot Interaction (HRI 2009)*.

Lohse, M., Hegel, F., Swadzba, A., Rohlfing, K., Wachsmuth, S., & Wrede, B. (2007). What can I do for you? Appearance and Application of Robots. *Workshop on The Reign of*

Catz and Dogz? The role of virtual creatures in a computerised society. Newcastle upon Tyne, GB, 121-126.

Lohse, M., Hegel, F., & Wrede, B. (2008). Domestic Applications for social robots - a user study on appearance and function. *Journal of Physical Agents, 2*, 21-32.

Lohse, M., Rohlfing; K., Wrede, B., & Sagerer, G. (2008). Try Something Else! When Users Change Their Discursive Behavior in Human-Robot Interaction. *Proceedings of 2008 IEEE International Conference on Robotics and Automation,* Pasadena, CA.

Maas, J.F. (2007). Dynamische Themenerkennung in situierter Mensch-Roboter-Kommunikation [Dynamic topic recognition in situated human-robot communication]. Doctoral Thesis. Applied Informatics, Bielefeld University, Technical Faculty, Germany.

Maddux, J. (1999). Expectancies and the Social-Cognitive Perspective: Basic Principles, Processes, and Variables. In I. Kirsch (Ed.). *How expectancies shape experience*. Washington, DC: American Psychological Association, 17-39.

Magnusson, D. (1981a). Problems in Environmental Analysis - An Introduction. In D. Magnusson (Ed.). *Towards a Psychology of Situations*. Hillsdale, NJ: Lawrence Erlbaum Associates, 3-7.

Magnusson, D. (1981b). Wanted: A Psychology of Situations. In D. Magnusson (Ed.). *Towards a Psychology of Situations*. Hillsdale, NJ: Lawrence Erlbaum Associates, 9-32.

McNeill, D. (1992). *Hand and Mind: what gestures reveal about thought.* Chicago: University of Chicago Press.

McNeill, D. (2005). *Gesture & Thought*. Chicago, London: University of Chicago Press.

Mercer, N. (1992). Culture, context and the construction of knowledge in the classroom. In P. Light & G. Butterworth (Eds.). *Context and Cognition*. London: Prentice Hall, Harvester Wheatsheaf, 28-46.

Milgram, S. (1974). *Obedience to Authority: An Experimental View*. New York: Harper and Row.

Mills, G. & Healey, P. (2006). Clarifying spatial descriptions: Local and global effects on semantic co-ordination. *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue.* Potsdam. Germany.

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

Nagai, Y., Asada, M., & Hosoda, K. (2006). Learning for joint attention helped by functional development. *Advanced Robotics, 20 (10)*, 1165 – 1181.

Nehaniv, C., Dautenhahn, K., Kubacki, J., Haegele, M., Parlitz, C., & Alami, R. (2005). A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. *Proceedings IEEE*

International Conference on Robot & Human Interactive Communication (ROMAN'05), 371-377.

Neuberg, S. L. (1996). Social Motives and Expectancy-Tinged Social Interactions. In R. M. Sorrentino & E. T. Higgins (Eds.). *Handbook of Motivation and Cognition*. The Guilford Press: NewYork, London, *3,* 225-261.

Nielsen, J. (1993). *Usability Engineering*. San Francisco: Kaufmann.

Nomura, T., Kanda, T., & Suzuki, T. (2004). Experimental Investigation into Influence of Negative Attitudes toward Robots on Human-Robot Interaction. *Presented at the 3rd Workshop on Social Intelligence Design (SID2004),* Twente.

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2004). Psychology in Human Robot Communication: An Attempt through Investigation of Negative Attitudes and Anxiety toward Robots. *Proceedings IEEE International Conference on Robot & Human Interactive Communication (ROMAN'04)*.

Olson, J., Roese, N., & Zanna, M. (1996). Expectancies. In E. Higgins & A. Kruglanski (Eds.). *Social Psychology: Handbook of Basic Principles*. New York: Guilford Press, 211-238.

Oppermann, D., Schiel, F., Steininger, S., & Beringer, N. (2001). Off-Talk - a Problem for Human-Machine-Interaction. *Proc. Eurospeech01, Aalborg,* 2197-2200.

Otero, N., Nehaniv, C., Syrdal, D., & Dautenhahn, K. (2006). Naturally Occurring Gestures in a Human-Robot Teaching Scenario. *Proc. 15th IEEE Int Symposium on Robot and Human Interactice Communication (ROMAN'06)*, 533-540.

Pacchierotti, E.; Christensen, H. I. & Jensfelt, P. (2005). Human-Robot Embodied Interaction in Hallway Settings: a Pilot User Study. *ROMAN 2005,* 164-171.

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.

Pfeifer, R. & Scheier, C. (1999). *Understanding Intelligence.* Cambridge, MA: MIT Press.

Ray, C., Mondada, F., & Siegwart, R. (2008). What do people expect from robots?. *Proceedings 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems,* Nice, France, 3816-3821.

Remland, M. S., Jones, T. S., & Brinkman, H. (1995). Interpersonal Distance, Body Orientation, and Touch: Effects of Culture, Gender, and Age. *The Journal of Social Psychology, 135(3)*, 281-297.

Roese, N. J. & Sherman, J. W. (2007). Expectancy. In A. W. Kruglanski & E. T. Higgins (Eds.). *Social Psychology. Handbook of Basic Principles.* New York: Guilford Press, 91-115.

Rohlfing, K. (to appear). Meaning in the objects. In J. Meibauer & M. Steinbach (Eds.). *Experimental Pragmatics/Semantics.* Amsterdam: John Benjamins.

Rohlfing, K., Rehm, M., & Goecke, K. (2003). Situatedness: The interplay between context(s) and situation. *Journal of Cognition and Culture, 3(2)*, 132-157.

Ross, L. & Nisbett, R. E. (1991). *The person and the situation : perspectives of social psychology*. New York, London: McGraw-Hill.

Ross, M. & Conway, M. (1986). Remembering one's own past: The construction of personal histories. In R. M. Sorrentino, & E. T. Higgins, (Eds.). *The handbook of motivation and cognition: Foundations of social behavior*. New York: Guilford Press, 122-144.

Rotter, J. B. (1981). The Psychological Situation in Social-Learning Theory. In D. Magnusson (Ed.). *Towards a Psychology of Situations*. Hillsdale, NJ: Lawrence Erlbaum Associates, 169-178.

Sabanovic, S., Michalowski, M., & Caporael, L. (2007). Making Friends: Building Social Robots through Interdisciplinary Collaboration. *Multidisciplinary Collaboration for Socially Assistive Robotics: Papers from the 2007 AAAI Spring Symposium, Technical Report SS-07-07*, 71-77.

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

Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Schank, R. C. (1999). *Dynamic Memory Revisited*. Cambridge, MA: Cambridge University Press.

Schank, R. & Abelson, R. (1975). Scripts, Plans, and Knowledge. *In the Proceedings of the Fourth International Joint Conference on Artificial Intelligence, Tblisi, USSR,* 151-157.

Sears, D. O., Peplau, L. A., Freedman, J. L., & Taylor, S. E. (1988). *Social Psychology*. Englewood Cliffs, NJ: Prentice Hall.

Severinson-Eklundh, K., Green, A.; Hüttenrauch, H., Oestreicher, L., & Norman, M. (2003). Involving Users in the Design of a Mobile Office Robot. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 34(2),* 113-124.

Shepherd, A. (2001). *Hierarchical Task Analysis.* London and New York: Taylor and Francis.

Shneiderman, B. (2002). *User Interface Design*. Bonn: mitp.

Shuter, R. (1977). A Field Study of Nonverbal Communication in Germany, Italy, and the United States. *Communication Monographs, 44 (4),* 298-305.

Sidner, C. L., Kidd, C., Lee, C., & Lesh, N. (2004). Where to look: A study of human-robot engagement. *In Intelligent User Interfaces (IUI),* Funchal, Island of Madeira, Portugal, January 2004, 78-84.

Siegrist, J. (2005). *Medizinische Soziologie* [Medical Sociology]. München: Elsevier.

Smith, L. B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review, 25*, 278-298.

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

Staudte, M. & Crocker, M. (2008). The utility of gaze in spoken human-robot interaction. *Proceedings of Workshop on Metrics for Human-Robot Interaction 2008, March 12th,* Amsterdam*,* the Netherlands, 53-59.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common Metrics for Human-Robot Interaction. *The First ACM/IEEE International Conference on Human-Robot Interaction (HRI 2006).*

Suzuki, S., Morishima, Y., Nakamura, M., Tsukidate, N., & Takeda, H. (2007). Influence of body orientation and location of an embodied agent to a user. *Proceedings of the 20th International Conference on Computer Animation and Social Agents (CASA2007),* Hasselt University, Belgium*,* 1-10.

Syrdal, D. S., Koay, K. L., Walters, M. L., & Dautenhahn, K. (2007). A personalized robot companion ? The role of individual differences on spatial preferences in HRI scenarios. *Proceedings of the 16th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2007)*, 26-29.

Takeda, H., Kobayashi, N., Matsubara, Y., & Nishida, T. (1997). Towards Ubiquitous Human-Robot Interaction. *Proc. IJCAI Workshop on Intelligent Multimodal Systems,* Nagayo, Japan.

Taylor, S. &. Fiske, S. (1975). Point of View and Perception of Causality. *Journal of Personality and Social Psychology, 32*, 439-445.

Thrun, S. (2004). Towards a Framework for Human-Robot Interaction. *Human-Computer Interaction, 19 (1&2)*, 9-24.

Tolman, E. C. (1932). *Purposive behavior in animals and men.* New York: Appleton-Century.

Vollmer, A., Lohan, K. S., Fischer, K., Nagai, Y., Pitsch, K., Fritsch, J., Rohlfing, K. J., & Wrede, B. (2009). People Modify Their Tutoring Behavior in Robot-Directed Interaction for Action Learning. *Proceedings International Conference on Development and Learning 2009,* Shanghai, China.

Wagner, A. & Arkin, R. (2006). A Framework for Situation-based Social Interaction. *The 15th IEEE International Symposium on Robot and Human Interactive Communication, (RO-MAN'06).*

Wagner, A. R. & Arkin, R. C. (2008). Analyzing Social Situations for Human-Robot Interaction. *Human and Robot Interactive Communication: Special Issue of Interaction Studies, 9(2)*, 277-300.

Walters, M., Koay, K. L., Woods, S., Syrdal, D. S., & Dautenhahn, K. (2007). Robot to Human Approaches: Preliminary Results on Comfortable Distances and Preferences. *Technical Report of the AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics, (AAAI SS07-2007*), Stanford University, Palo Alto, CA, USA.

Wilson, T. D. & Hodges, S. D. (1992). Attitudes as temporary constructions. In L.L. Martin & A. Tesser (Eds.). *The construction of social judgments.* Hillsdale, NJ: Lawrence Erlbaum, 37-66.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* [Interrater agreement and interrater reliability]. Göttingen: Hogrefe. Verlag für Psychologie.

Wrede, B., Kopp, S., Rohlfing, K. J., Lohse, M., & Muhl, C. (to appear): Appropriate feedback in asymmetric interactions. *Journal of Pragmatics*.

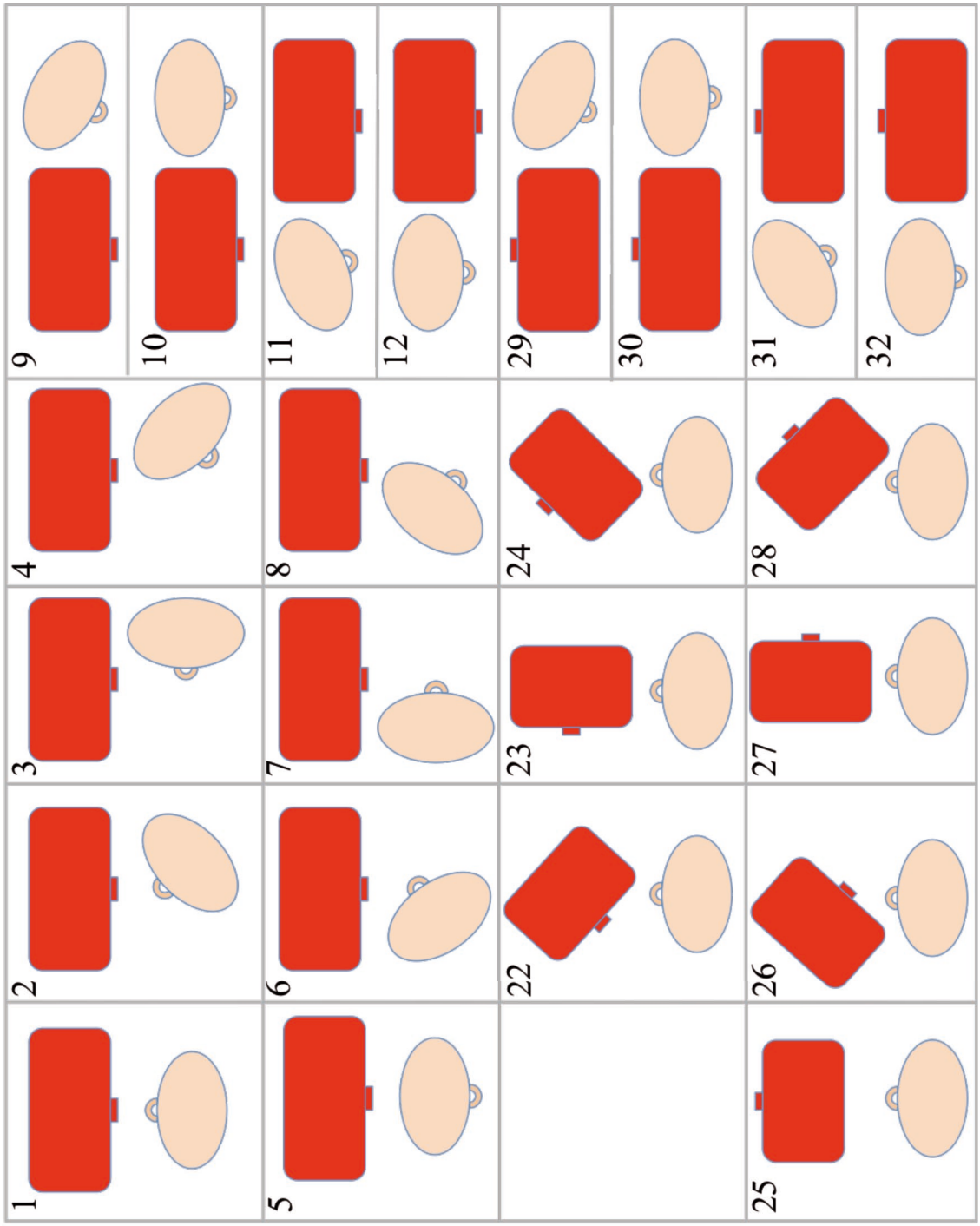
Yamazaki, A., Yamazaki, K., Kuno, Y., Burdelski, M., Kawashima, M., & Kuzuoka, H. (2008). Precision timing in human-robot interaction: Coordination of head movement and utterance. *Proceedings CHI 2008,* 131-140.

Young, J. E., Hawkins, R., Sharlin, E., & Igarashi, T. (2009). Toward Acceptable Domestic Robots: Applying Insights from Social Psychology. *International Journal of Social Robotics, 1,* 95-108.

Zimbardo, P. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. New York: Random House.

Appendix

Appendix A – coding scheme body orientation



Appendix B – interpretation predecessor/successor transition matrix

While most results of the SALEM are self-explaining, the interpretation of the transition matrixes might not be. Therefore, in the following their output is briefly explained. Assume that the output of the predecessor transition matrix looks as follows for three different annotation values (1, 2, 3):

Table 0-1. Example of predecessor transition matrix

value	1	2	3
1	0	0.8545	0.1455
2	0.8367	0	0.1633
3	0.8824	0.1176	0

In the case depicted here, if the value of the current annotation is '1' (first row), the probability that the previous annotation had the value '2' is 0.8545 (85.45%). In the predecessor matrix the rows sum up to 1 (or 100%). In contrast, in the successor matrix the columns sum up to 1 (100%). The following table again shows an example for three different annotation values (1, 2, 3):

Table 0-2. Example of the successor transition matrix

value	1	2	3
1	0	0.8790	0.4712
2	0.4867	0	0.5288
3	0.5133	0.1210	0

In this example, the probability that the annotation value '1' (first column) is succeeded by value '2' is .4867 (48.67%).

Appendix C – questionnaire home tour studies (results for the first and the second iteration)

	mean
How much do you like BIRON? (1 not at all … 5 very much)	4.01
How satisfied are you with BIRON's behavior? (1 not at all … 5 very much)	3.22
Please rate BIRON on the following scales:	
intelligent (1 stupid … 5 intelligent)	2.67
predictable (1unpredictable … 5 predictable)	3.34
consistent (1 inconsistent … 5 consistent)	3.59
talkative (1 quiet … 5 talkative)	3.20
fast (1 slow … 5 fast)	1.95
interested (1 not interested … 5 interested)	3.70
active (1 not active … 5 active)	3.26
polite (1 impolite … 5 polite)	4.67
friendly (1 unfriendly… 5 friendly)	4.63
obedient (1 not obedient … 5 obedient)	4.10
interesting (1 boring … 5 interesting)	2.63
useful (1 useless … 5 useful)	3.14
attentive (1 not attentive … 5 attentive)	4.09
practical (1 impractical … 5 practical)	2.54
cooperative (1 not cooperative … 5 cooperative)	4.00
funny (1 serious … 5 funny)	3.46
Questions for degree of agreement (1 don’t agree … 5 completely agree)	
I can hear what BIRON says well	4.80
BIRON understands me correctly	2.97
The conversation with BIRON is fluent	2.50
The interaction with BIRON is hard	3.46
The interaction with BIRON is easy to learn	4.25
The interaction with BIRON is frustrating	2.45
It is easy to get BIRON to do what I want	3.13
The interaction with BIRON is not flexible enough	3.34
A high degree of concentration is needed to handle BIRON	3.42
Overall it is easy to handle BIRON	3.76

Appendix D – guided question for the interviews in the home tour studies

1. Wenn Sie über Ihre speziellen Eindrücke nachdenken – wie war die Interaktion mit BIRON? (If you think about your impressions - how was the interaction with BIRON?)

2. Was ist Ihnen besonders aufgefallen? positiv oder negativ? Gab es besondere Situationen? (What did stand out? positively or negatively? Were there any special situations?)

3. Hat die Spracheingabe für Sie funktioniert? (Did the speech input work out for you?)

4. Fanden Sie die Sprachausgabe von BIRON angemessen und verständlich? (Did you find BIRON's speech output appropriate and comprehensible?)

5. Worauf haben Sie noch bei BIRON geachtet? Wie wussten Sie was der Roboter gerade macht? (What about BIRON did you pay attention to? How did you know what the robot was doing?)

6. Haben Sie eventuell weitere Anmerkungen oder Kommentare zur Studie oder BIRON? Hat der Versuch Spass gemacht? (Do you have anymore comments about the study or BIRON? Did you have fun?)

Appendix E – instructions for object-teaching study 1

[please turn to the next page for the English translation]

Auf diesem Bild sehen sie BIRON. Bei der Entwicklung des Roboters streben wir einen Assistenten im Haushalt an, das heißt, er soll in der Lage sein, sich in einer Wohnung zu orientieren und verschiedene Aufgaben des täglichen Lebens zu erledigen.
Um sich in seiner Umgebung zurechtzufinden, muss der Roboter lernen können. Ziel des Experiments ist es deshalb, BIRON verschiedene Objekte zu zeigen. Sie werden während der Interaktion mit dem Roboter eine Auswahl neben sich auf einem Tisch finden. Bitte beachten Sie, dass BIRON die Objekte vor der Interaktion mit Ihnen nicht kennt und erst lernen muss. Bitte zeigen Sie BIRON ungefähr 5 Objekte und bringen Sie ihm deren Namen bei.

Wichtig beim Gespräch mit BIRON:

1. Bitte bedenken Sie, dass der Roboter manchmal etwas mehr Zeit für die Verarbeitung braucht .
2. Bitte versuchen Sie während des Experiments alle Probleme mit dem Roboter und nicht mit den anderen anwesenden Personen zu lösen.
3. Damit BIRON Sie nicht "aus den Augen verliert", ist es wichtig, dass Sie möglichst die ganze Zeit mit beiden Beinen nebeneinander fest auf dem Boden stehen.

Der Ablauf des Gesprächs sollte darüber hinaus, wie mit einem Menschen auch erfolgen (begrüßen sie BIRON, unterhalten Sie sich mit ihm, verabschieden Sie sich von ihm). Sie haben ungefähr 10 Minuten Zeit sich mit BIRON zu unterhalten.

Viel Spaß!

This is BIRON. We strive to develop an assistant for the household which means that the robot shall be able to find its way around in an apartment and to solve various everyday-life tasks.
To get along in its surroundings, the robot has to be able to learn. Therefore, the goal of the experiment is to teach different objects to BIRON. During the interaction with the robot, you will find a selection on a table next to you. Please keep in mind that BIRON does not know the objects before the interaction with you and needs to learn them. Please show about 5 objects to BIRON and teach it their names.

Important when talking to BIRON:

1. Please keep in mind that the robot sometimes needs some time for computation.
2. During the experiment, please try to solve all upcoming problems with the robot and not with other people who are present.
3. It is important that you stand with both feet on the ground whenever possible that BIRON does not lose sight of you.

Apart from that, the interaction should be like with a human (greet BIRON, talk to it, say goodbye). You have about ten minutes to talk to BIRON.

Have fun!

Appendix F – instructions for object-teaching study 2

Auf diesem Bild sehen sie BIRON. Bei der Entwicklung des Roboters streben wir einen Assistenten im Haushalt an, das heißt, er soll in der Lage sein, sich in einer Wohnung zu orientieren und verschiedene Aufgaben des täglichen Lebens zu erledigen.
Um sich in seiner Umgebung zurechtzufinden, muss der Roboter lernen können. Ziel des Experiments ist es deshalb, BIRON verschiedene Objekte zu zeigen. Sie werden während der Interaktion mit dem Roboter eine Auswahl neben sich auf einem Tisch finden. Bitte beachten Sie, dass BIRON die Objekte vor der Interaktion mit Ihnen nicht kennt und sie erst lernen muss. Bitte zeigen Sie BIRON nacheinander einige Objekte und bringen Sie ihm deren Namen bei. Bitte überprüfen Sie auch, ob er sie tatsächlich gelernt hat.

Wichtig beim Gespräch mit BIRON:

1. Bitte bedenken Sie, dass der Roboter manchmal etwas mehr Zeit für die Verarbeitung braucht .
2. Bitte versuchen Sie während des Experiments alle Probleme mit dem Roboter und nicht mit den anderen anwesenden Personen zu lösen.
3. Damit BIRON und die Kameras Sie nicht "aus den Augen verlieren", ist es wichtig, dass Sie möglichst die ganze Zeit unmittelbar vor dem Tisch stehen.

Der Ablauf des Gesprächs sollte darüber hinaus wie mit einem Menschen auch erfolgen (begrüßen Sie BIRON, unterhalten Sie sich mit ihm, verabschieden Sie sich von ihm). Bitte führen Sie das Experiment in zwei Phasen durch, die Objekte werden zwischendurch einmal ausgetauscht. Sie haben für jede Phase ungefähr 10 Minuten Zeit. Nach dem Experiment bitten wir Sie noch einen kurzen Fragebogen auszufüllen.
Sie werden während des Experiments gefilmt. Sie können das Experiment jederzeit ohne Angabe von Gründen abbrechen oder Teile bzw. Fragen auslassen.
Viel Spaß!

This is BIRON. We strive to develop an assistant for the household which means that the robot shall be able to find its way around in an apartment and to solve various everyday-life tasks.
To get along in its surroundings, the robot has to be able to learn. Therefore, the goal of the experiment is to teach different objects to BIRON. During the interaction with the robot, you will find a selection on a table next to you. Please keep in mind that BIRON does not know the objects before the interaction with you and needs to learn them. Please show some objects, one after the other, to BIRON and teach it their names. Please make sure that it has learned them.

Important when talking to BIRON:

1. Please keep in mind that the robot sometimes needs some time for computation.
2. During the experiment, please try to solve all upcoming problems with the robot and not with other people who are present.
3. It is important that you stand right in front of the table the whole time that BIRON does not lose sight of you.

Apart from that, the interaction should be like with a human (greet BIRON, talk to it, say goodbye). The experiment has two phases between which the objects will be exchanged. You have about ten minutes to talk to BIRON in each phase. After the experiment we will ask you to answer a short questionnaire.
You will be videotaped during the experiment. You can abort the experiment or skip parts or questions without providing reasons whenever you like.

Have fun!

Appendix G – instructions for the home tour studies

Anweisungen zum Versuch mit BIRON

Willkommen zu unserem Versuch. Im Folgenden erhalten Sie wichtige Informationen und Anweisungen.

Das Szenario:
Stellen Sie sich vor, Sie haben einen Roboter gekauft, der Ihnen im Haushalt helfen soll. Wenn Sie den Roboter zu Hause ausgepackt haben, müssen Sie ihm zuerst Ihre Wohnung zeigen. Der Roboter muss dabei lernen, wie die Räume heißen und wo sich wichtige Gegenstände befinden.

Der Roboter BIRON:
BIRON ist ein Prototyp, mit dem wir das eben beschriebene Szenario untersuchen. Er kann Spracheingaben auf Deutsch verstehen und darauf reagieren. BIRON sieht, dass ein Mensch anwesend ist und kann diesem folgen. Außerdem kann er sich die Namen von Räumen und Gegenständen merken.

Ihre Aufgabe:
Der Versuch ist in drei Abschnitte unterteilt:

1. sich mit der Spracheingabe vertraut machen
2. sich mit BIRON vertraut machen, den Roboter begrüßen und fragen, was er kann
3. mit BIRON verschiedene Räume besuchen und dem Roboter Gegenstände zeigen

---

Teil 1: Machen Sie sich mit der Spracheingabe vertraut
Sie bekommen von uns gleich ein tragbares Mikrofon. Dieses hilft BIRON zu hören, was Sie sagen. Ihre Aufgabe ist es, einige einfache Sätze zu sprechen, damit wir überprüfen können, ob die Spracheingabe funktioniert.

Teil 2: Machen Sie sich mit BIRON vertraut

BIRON kann nur einfache Sätze verarbeiten. Im Folgenden finden Sie einige Sätze, die der Roboter gut verstehen kann. Diese sollen Ihnen als Hilfestellung dienen. Sie dürfen aber auch gern andere Sätze verwenden, da der Roboter nicht nur die unten genannten versteht.

Begrüßung

„Hallo BIRON“

Zeigen und Benennen eines Raumes bzw. eines Gegenstandes

„BIRON das ist das Wohnzimmer“

Wenn ein Raum benannt wird, schaut sich BIRON um, um diesen zu erlernen.

„BIRON das ist ein Regal“

Bitte zeigen Sie auf Gegenstände und benennen Sie diese, wenn BIRON sie lernen soll.

BIRON durch die Wohnung führen

Bitten Sie BIRON, Ihnen zu folgen, um zusammen mit ihm in einen anderen Raum zu gehen.

„BIRON, folge mir.“

BIRON helfen

BIRON braucht ab und zu Ihre Hilfe, z.B. wenn er nicht durch die Tür kommt oder ein Hindernis sieht. Der Roboter sagt dann z.B.:

> „Ich vermute, dass ich von einem Hindernis behindert werde. Bitte schiebe mich ein wenig davon weg und sage noch einmal „hallo“.“

Wir zeigen Ihnen, wo und wie Sie BIRON anfassen können, um den Roboter manuell zu bewegen. Damit können Sie ihm helfen, durch die Tür zu gehen oder sich von Hindernissen zu entfernen. Bitte sagen Sie danach noch einmal „Hallo“ zu BIRON.

BIRON stoppen

Sie können BIRON jederzeit anhalten.

„BIRON anhalten“ bzw. „BIRON stopp“

BIRON neu starten

Es kann vorkommen, dass BIRON scheinbar nicht mehr reagiert. Wenn Sie das Gefühl haben, dass Sie nicht mehr weiterkommen, können Sie den Roboter neu starten.

„BIRON Neustart“

BIRON startet neu und erwartet danach wieder eine Begrüßung (z.B. „Hallo“) von Ihnen.

Von BIRON verabschieden

Wenn Sie die Interaktion mit BIRON unterbrechen oder beenden wollen, können Sie sich verabschieden.

„BIRON tschüss“

Wenn BIRON piept

BIRON gibt manchmal kurze Pieptöne von sich, speziell beim Losfahren. Diese Töne zeigen eine Motorfunktion an und haben keine Bedeutung für die Interaktion.

Teil 3: Zeigen Sie BIRON die Wohnung

Sind Sie bereit, BIRON die Wohnung zu zeigen?

First iteration:

- Führen Sie BIRON von der Batterieladestation zur Mitte des Raumes und zeigen Sie ihm das Wohnzimmer.
- Führen Sie BIRON zu einem grünen Sessel und zeigen Sie ihm diesen Gegenstand.
- Fahren Sie mit BIRON in das Esszimmer und zeigen Sie dem Roboter diesen Raum.
- Führen Sie BIRON zum Tisch im Esszimmer. Zeigen Sie dem Roboter den Tisch.
- Verabschieden Sie sich von BIRON.

Second iteration:

- Führen Sie BIRON von der Batterieladestation zur Mitte des Raumes und zeigen Sie ihm das Wohnzimmer.
- Zeigen Sie BIRON eines der Regale.
- Fahren Sie mit BIRON in das Esszimmer und zeigen Sie dem Roboter diesen Raum.
- Führen Sie BIRON zum Tisch im Esszimmer. Zeigen Sie dem Roboter das darauf liegende Buch.
- Verabschieden Sie sich von BIRON.

Instructions for the interaction with BIRON

Welcome to our study. In the following you receive important information and instructions.

The scenario:
Imagine that you have bought a robot to help you in the household. When you have unpacked the robot at home you first need to show it the apartment. The robot has to learn how the rooms are called and where important objects are located.

The robot BIRON:
BIRON is a prototype with which we investigate the scenario described above. It can understand speech input in German and react to it. BIRON sees that a human is present and can follow him or her. Moreover, it can memorize the names of objects and rooms.

Your task:
The study consists of three tasks:

1. acquaint yourselve with the speech input system
2. acquaint yourselve with BIRON, greet the robot and ask what it can do
3. visit different rooms with BIRON and show objects to the robot

---

Part 1: acquaint yourselve with the speech input system
In a moment you will get a portable microphone. This helps BIRON to hear what you say. Your task is to speak some simple sentences which helps us to test if the speech input system works.

Part 2: acquaint yourselve with BIRON

BIRON can only compute simple sentences. In the following, you find some sentences that the robot can understand well. These shall serve you as an aid. However, you are free to use other utterances because the robot does not only understand the ones mentioned below.

Greeting

"Hello BIRON"

Showing and naming rooms and objects

"BIRON this is the living room"

If a room is named, BIRON turns around to learn it.

"BIRON this is the shelve"

Please point at objects and name them in order for BIRON to learn them.

Guiding BIRON through the apartment

Ask BIRON to follow you to go to another room together.

"BIRON, follow me."

Helping BIRON

Once in a while BIRON needs your help, for example, when it cannot cross the door or sees an obstacle. The robot then says, for example:

„I think that there is an obstacle in my way. Please pull me away from it and say hello again."

We show you how to touch BIRON in order to move it manually. Therewith you can help it to cross the door or to move away from obstacles. Please say "hello" again to BIRON afterwards.

Stopping BIRON

You can stop BIRON anytime.

“BIRON stop“ or „BIRON stop“

Resetting BIRON

It may happen that BIRON does not seem to react anymore. If you feel that you cannot advance, you can reset the robot.

“BIRON reset“

BIRON restarts and expects a greeting of you (for example, “hello“).

Saying goodbye to BIRON

If you like to interrupt or finish the interaction with BIRON you can say goodbye.

“BIRON bye“

If BIRON beeps

Sometimes BIRON beeps, especially when starting to drive. The beeps indicate a motor function and do not have any meaning for the interaction.

Part 3: Show the apartment to BIRON

Are you ready to show the apartment to BIRON?

First iteration:

- Guide BIRON from the charging station to the middle of the room and show it the living room
- Guide BIRON to the green armchair and show it this object
- Drive BIRON to the dining room and show it this room
- Guide BIRON to the table in the dining room. Show the table to the robot
- Say goodbye to BIRON

Second iteration:

- Guide BIRON from the charging station to the middle of the room and show it the living room
- Show one of the shelves to BIRON
- Drive BIRON to the dining room and show it this room
- Show one of the floor lamps to BIRON
- Say goodbye to BIRON