FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D1.3.1

Report on the evaluation methodology of multi-modal dialogue

Due date of deliverable: 31/12/2004
Actual submission date: 31/01/2005

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

Organisation name of lead contractor for this deliverable:
KTH Numerical Analysis and Computer science

Revision: Final
Dissemination Level: PU

Executive Summary

The purpose of this report is to survey and to describe the way hi-fi simulation methods are used to evaluate project specific dialogue models. These dialogue models have been developed within the Home-Tour Scenario (Key Experiment 1) by the University of Bielefeld. Using the dialogue models we have set up an environment in the CAS "Living room", equipped with furniture and objects that may be found in a normal living room. The environment is used as a context for a simulation study in the Wizard-of-Oz framework, meaning that the dialogue and actions of the robot are simulated by human operators, without informing the user about the the real circumstances of the study, until the end of the study. This allows for evaluation of what the user believes to be a working system.

We have collected data from four pilot sessions and 22 user sessions (lasting $\sim$ 15 minutes). The data amounts to about 5.5 hours of digital video and internal audio. We also collected images from four network cameras placed in each corner of the experiment room. On-board data from the laser range finder was also collected during the sessions. At this point we have transcribed and made preliminary analyses of the first five sessions in order to explore different ways of analysing the whole data set. Besides the linguistic information we get from the study, we also observed a set of interaction patterns that we found were of interest in the future study of the scenario. These patterns range from momentaneous requests for the robot's attention to patterns of deictic gestures used to reference objects (e.g. focusing) to elaborate ways of probing the robot's capability.

In the future we will continue our efforts to analyse the data that was collected in the study, focusing on development of tools and theory-based analysis models.

The role of evaluation in COGNIRON

We see the development of the cognitive robot companion as a stepwise process. By allowing users to interact with prototypes of different kinds we are able to get an understanding of the way humans might engage in interaction with future service robots.

A service robot with cognitive capabilities will in a specific and limited way, become part of a social context. This means that we need to investigate how the robot may go about to form new social relations with people. From a communicative point of view, the development of different styles of interaction is a result of agents trying to conduct some activity using their communicative skills. Our assumption is *not* that a robot will take part in such a process by emulating a human agent. Instead the robot will have limitations, but also capabilities that allow it to take on a new role in the co-creation of new specialised human-robot sub-languages.

Each component developed within COGNIRON will most certainly be evaluated according to the standards that are applicable to that specific component. However, in order to evaluate a complete cognitive companion we need to use prototypes to be able to communicate the design decisions to the user. By performing user studies during the early stages of the development process we are able to influence the design of the whole system. Furthermore we will gain a general understanding of multi-modal human-robot communication. The future goal is to be able to create guidelines for the design of communicative human-robot interfaces, influenced by theories of human-human communication, supported by observations from human-robot communication.

Relation to the Key Experiments

The evaluation study described in this report is primarily related to Key Experiment 1, the “Home Tour” (KE1). We are evaluating a system that is based on dialogue patterns developed by University of Bielefeld, allowing the user to show objects and locations to the robot. However, since the study allows for interaction with what appears to the user as a complete system, we anticipate results to have a bearing also general aspects related to human robot communication KE2 (the Curious robot) and KE3 (Learning and imitation).

Contents

Evaluation methodology of multi-modal dialogue

Anders Green, Helge Hüttenrauch and Kerstin Severinson Eklundh
{green,hehu,kse}@nada.kth.se

1 Evaluation of interface design for multi-modal human-robot dialogue

1.1 Introduction

Developing a complex system such as a service robot in a research context, means that iteration cycles are long ($\sim$ 1-2 years). In the COGNIRON project we are in the middle of the first iteration of development leading to an early working prototype or a demonstrator that could be tested with naïve but cooperative users. The purpose of this report is to present the methods used for evaluation of the particular dialogue design chosen for the robot. Our aim is to portray our interactive system in such a way that the users engage in interaction as they would with a real system. In this way we will gain an increased understanding of what behaviour we may expect from users of future service robots.

In this report we will describe and discuss the methods that may be used during the early stages of the design process, i.e., where there is still no working prototype, but only simulated or low-fi prototypes. We will also provide some preliminary results from the hi-fi simulation study that has been carried out during the first year of the project.

1.2 Iterative development of human-robot interfaces

The creation of an interface for a cognitively inspired robot is to some extent a very different undertaking than developing a graphical user interface for software application for the standard desktop computer. However, given the fact that we in a sense are developing a “product”, we may still employ methods from user-centred software engineering. This involves using a type of process oriented work flow similar to what is an established practice in more straight forward product development. Since human-robot can still can be regarded as pristine territory from a research point-of-view, we may be eclectic about the selection of methods that we employ in this process. We may therefore use already well-established methods together with novel techniques specific to human-robot interaction.

Normally a product that is released to the market has undergone several iterations of design and re-design. This process may be of lesser or greater complexity and the methods and work practices of may differ. However, viewing the development as a process is widely accepted. The development process for a human-robot interfaces can be viewed as a process like Figure 1. For more clear cut development of speech interfaces, Hulstijn (2000) identified activities of four main types leading towards the complete and working system: *analysis, design, implementation and evaluation.*. Along the time-line, these activities are iterated. To this picture we may add a set of stake-holders: users, designers, developers, etc, who have an interest or will be affected by the design of the system.

This paper describes methods related to our ongoing work in evaluating prototypes for multi-modal human robot dialogue. We are therefore limiting our effort to giving a brief overview of methods that

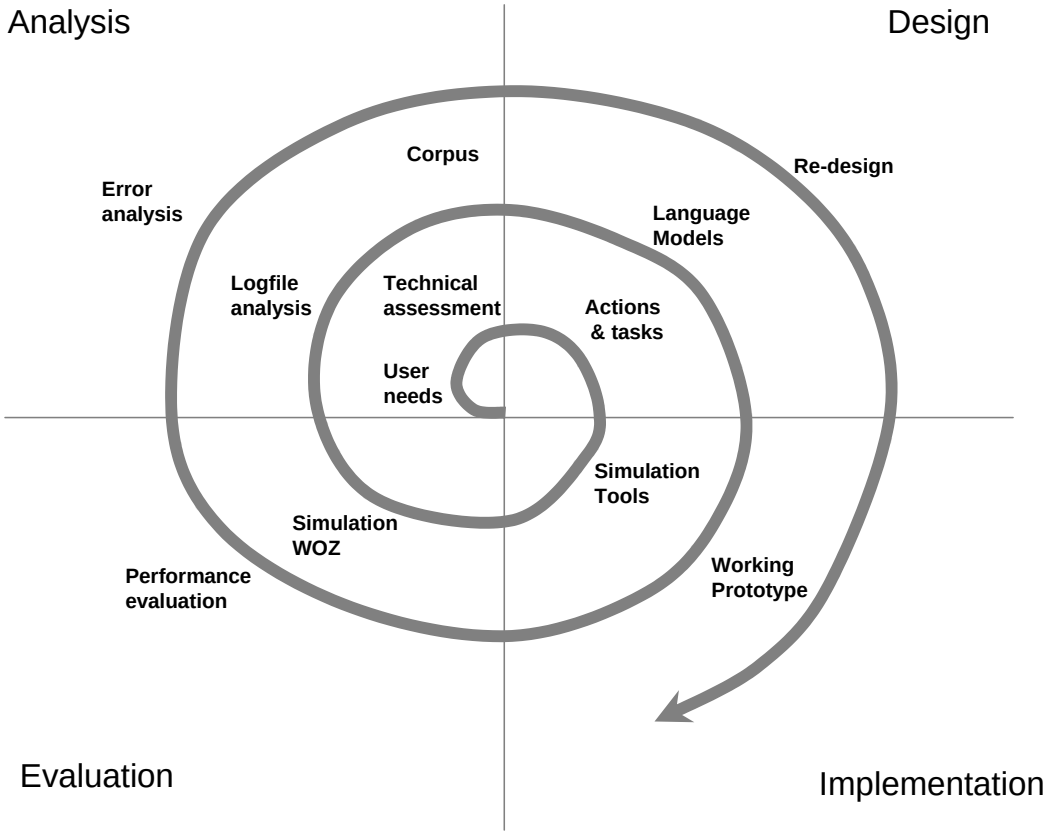


Figure 1: Iterative development cycle, (modelled after Hulstijn, 2000)

we can relate to this stage of development. When creating a system that does not yet exist, like a service robot, we need methods that allow us to evaluate different kinds of interface metaphors before we put large effort into building a working prototype. Thus the focus of this overview is concentrated on methods for collecting data on users' attitudes and behaviour: Rapid prototyping, Focus groups, Verbal protocols and Hi-fi simulation techniques.

Given the early stage of the development process we will not discuss methods that are primarily used for usability evaluation. Hence we will not discuss benchmarking and performance metrics (e.g. like Scholtz, 2002; Crandall & Goodrich, 2003) due to the explorative nature of the research presented in the following. Another area which we cannot really discuss is established guidelines specific to human-robot interaction. Very few attempts have been made to establish guidelines for *multi-modal interfaces* for service robots. This stems from the fact that human-robot interaction is a relatively new discipline lacking cases of actual product development aimed at end users. Instead our research aims to further develop our ability to create future design guidelines for the service robot domain in a similar fashion as for speech interfaces and multi-modal interfaces (e.g. Cheepen, Gilbert, Failenschmid, & Williams, 2002; Reeves et al., 2004).

1.3 Rapid prototyping

During the early phases of the system development methods that can be used to communicate the intended use of the product, e.g, sketching (Lövgren, 2004), mock-ups (Ehn & Kyng, 1992) and scenarios with Synthetic dialogues (Green & Severinson Eklundh, 2001) are of great interest in order to inform design. The "rapidness" is related to what may be seen as being rapid given the type of product. Thus card board models of design elements, a sketch of a scenario together with created dialogues can be put together in a short time. Video prototyping on the other hand may take considerable time, but may still be an early step, given the size of the project.

1.4 Focus groups

An informal technique that is widely spread is the use of focus groups to collect qualitative data to gain an understanding of the attitudes and background knowledge of users. A focus group is typically set up as a group interview with 5-10 participants selected based on the characteristics they share. A test leader facilitates and controls the flow of the discussion around a set of relevant questions.

Depending on the type of users that get involved in the focus groups results may also come in the form of precise requirements or design ideas. Here we find that that Wizard-of-Oz simulations and demos may be used to communicate design ideas to users in order to start discussions in focus groups (e.g. Green, Hüttenrauch, Norman, Oestreicher, & Severinson Eklundh, 2000; Mival, Cringean, & Benyon, 2004).

1.5 Verbal protocols

Torrance (1994) used a method for assessing users' conception of robot oriented dialogues, i.e., dialogues created by introspection about users own performance. Torrance asked some users to write down what they would say to the robot. He also asked them to rank these sentences in order of difficulty as perceived by the users. A common approach in usability assessment for development of graphical user interfaces is think-aloud protocols, where the user is told to verbalise his or her actions. In development of speech interfaces this method is not commonly used for obvious reasons. A related technology, post-verbalisation, has been proposed and used by (Karsenty, 2001). Instead of a continually commenting on his own performance, the user is prompted to comment on the systems performance or to formulate an utterance that he find appropriate at that point in interaction. This technology has been used in a slightly different variant by (James, Rayner, & Hockey, 2000). By using pre-recorded scenarios the user was able to view a screen visualization of the system and the spoken dialogue. During some points (typically in the end) the user was prompted to complete the dialogue; i.e. to say what the robot should do next; or to rate the dialogue using some evaluation measure.

1.6 Hi-fi simulation

We have already mentioned low-fi prototyping as a means of provide a focal point for reflection of both users and designers. When it comes to bring users into the design process and to communicate the conception of a natural language interface, i.e. the type of interface we foresee a cognitively inspired service robot to have, simulation techniques provide the closest thing to assessing interaction

with the real thing provided that the level of fidelity of the simulation is sufficient to create the illusion of interacting with a real system.

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

The starting point for a Wizard-of-Oz study involves the construction of a prototype where some features of the system are for real and where some functions are simulated by one or more operators who control the system's actions and responses. A classical setup is to put a user in front of a desktop computer in one room and an operator, a wizard, in another room. The user is given a scenario by a test leader; a set of tasks to solve using the novel system and the interactions between the user and the system are recorded. Since the user often is unacquainted with systems of that particular kind, e.g. speech interfaces, or the task, the characteristics of the setup are that of a kind of a role-play, where the user tries to engage and act within the given scenario. Once the experiment is started the user is allowed to interact with the system in the same way as if the system was for real. The wizard acts as the system's high-level reasoning component responding to the users actions. During the experiment the test leader may intervene if the user gets into trouble related to the use of the (simulated) system. After the user has completed the scenario the test leader normally performs some kind of post session interview. At the end of this interview the test leader briefs the user that the system was in fact simulated and asks permission to use the data that has been collected during the experiment for research purposes. If the user does not give his/her permission the recordings are noted as unanswered and the data media is erased. If there is some kind of reward to be given to the user this will be handed over irrespectively of a negative or positive reply to the request to use data.

Wizard-of-Oz simulation methodology in its classical form, i.e., where one user interacts with one (desktop) computer in a lab environment, has been used since the 1970s. The term itself was first used by Kelley (1984) two decades ago. Malhotra (1975) used the method for simulation of natural language expert systems. The method has also been used for simulation of database question answering systems (Dahlbäck, Jönsson, & Ahrenberg, 1993; Maulsby, Greenberg, & Mander, 1993) and unimodal speech interfaces (Antoniol, Cattoni, Cettolo, & Federico, 1993), but also for development of multi-modal interfaces (Oviatt, 1996; Salber & Coutaz, 1993; Perzanowski, Schultz, Adams, Marsh, & Bugajska, 2001).

One argument for extracting data from recorded human-human conversation instead of performing Wizard-of-Oz simulation is that we may capture behaviour of people bringing real tasks to the system as opposed to the role-play occurring in simulation (Jönsson & Dahlbäck, 2000).

In the Human-Robot Interaction domain there are some compelling reasons for performing simulations rather extracting data from human-human communication. In many of the envisioned scenarios human-robot communication is focused on communication about topics that are already known to the participants during conversation, making it unnecessary to address these explicitly using speech during the execution of a task.

First of all there is the matter of the task that the robot is performing. The tasks that we address the robot with are either simple or focused on domains where explicit verbal instructions are rare or only sparsely used by humans, but seem very important for robots. For instance, explicit negotiation of easily understood tasks (e.g. fetching an object from a known location) (Green, 2001), and communication concerning detection of humans and description of routes (Fry, Asoh, & Matsui, 1998).

Secondly, even in scenarios where robots are intended to replace people, in dirty, dangerous or distant environments, the work allocation between actual workers and robots may differ (see Casper & Murphy, 2003, for an example of extreme use). This means that the conversation between field workers, like divers or rescue workers would probably be of little direct use to when developing modules for spoken human-robot communication.

1.6.1 Simulation studies of Human-Robot Interaction

When mobility of both users and systems become a topic for investigation, the complexity of setting up the scenario increases, especially the amount of people required to maintain and control the scenario. However, when carefully designed, a simulation study will provide data about different aspects of human-robot interaction that would otherwise be unattainable until large effort had been spent on the creation of a working prototype. Thus, the data that we get from a Wizard-of-Oz study is to a large extent qualitative, but we may also collect data as a resource to be used for component development, e.g. as training data for speech recognisers. We may also use the collected data as for quantitative evaluation.

A thematic view of the type of data that has attracted the interest of researcher yields the following categories:

- Data on language use, especially spatial language (including gesture and speech).
- Visualisation of the whole interaction to enable the designer to conceptualise what the system could or should do in different realistic situations.
- Assessment of users' attitudes towards a future system or towards robots in general.

Language use

Compared to Wizard-of-Oz studies in uni-modal natural language interfaces, there are only a few examples of simulation studies aimed at human-robot interaction. The MAIA project employed this method to collect a corpus for a service robot (Corazza, Federico, Gretter, & Lazzari, 1993). The work was done assuming some prerequisites proposed by Fraser and Gilbert (1991):

- the future system is exactly specified.
- the future system is imitated realistically.
- the simulation is convincing

Corazza et al (1993) point out that it is not trivial to meet the first criterion in a robotics scenario and consequently they used an on screen visualisation where all the system components were simulated. When recording the user, a rigorous setup with an acoustically isolated room was used in order to be able to employ the corpus as training data for the automatic speech recogniser (ASR). This also allowed for quantitative testing of the ASR against the recorded data (Antoniol et al., 1993).

There are some interesting studies that to a large extent are aimed at trying out and conceptualising an interface for a service robot.

Perzanowski et al (2003) used a Wizard-of-Oz setup to collect data for spatial navigation using a multi-modal interface for the robot Coyote. The pilot study was the first step towards performing a more formal study. The multi-modal user interface involved a touch screen based graphical user interface with integrated speech input. The users were asked to find an object (the "foo") in another room. The analysis concerned the collected linguistic information (e.g. utterances, utterances coordinated with gestures) but the goal of the study was also to validate the Wizard-of-Oz method for a planned future study.

As one of the first steps in the user centred approach taken by Green et al (2000) a fairly unrestricted Wizard-of-Oz study was set up to explore the natural language interaction with a fetch and carry robot. The overall goal for the study was to illustrate a human-robot interaction scenario in an explorative manner in order to inform design of a robot for a service task in an office. In the study the users were instructed to use the robot to deliver and carry objects in to different place in a room where locations were defined to represent places in an office environment. In the trial the user was instructed to perform two different missions using the robot. Firstly the user should deliver an object (a magazine) using the robot. The user was instructed to stay and get the robot to fetch the remaining magazines. The second mission was to fetch a glass of water using the robot. The user was informed that the robot could accompany the user using a behaviour for person following. One of the most important results of the study was that users need continuous feedback of several kinds. The participants in the study were often confused about the robot's state, and closely monitored its movements. A multi-modal style of interaction (e.g. with pointing gestures) was used, even if users knew that the robot had no vision capabilities.

Spatial language

The focus of Perzanowski et al (2003) was apart from evaluation of the Wizard-of-Oz technique to collect data on spatial language (speech and gestures) used to manoeuvre a robot.

Data collection for spatial language is is one of the most frequent ways of using simulation techniques. For instance, Lauria et al (2002) collected route descriptions for robot navigation similar to the Map task corpus (Carletta et al., 1997). The practical setup in the collection was not a proper wizard setup but users were told that they were interacting with a human operator hidden in another room. The small robot used in the setup was located in a physical model world, with the user standing beside it.

The users were instructed to address the operator, sitting in another room, as if the operator could see the video image from a camera mounted on the robot. They were also told to re-use route information that had already been specified whenever possible. Using the corpus a grammar was constructed, specifying the primitive procedures used by the participants to describe the route for the robot.

Attitudes towards service robots

Sakamoto et al (2004) used a Wizard-of-Oz type of setup to measure attitudes (e.g. "Sharedness" and "Emphaty") related to "cooperative" body movements of a robot. The task for the user was to describe a route to the robot. During the interaction the robot issued gestures but remained in a fixed position in the room. Since the focus was on the gestures, the verbal output from the robot was limited. This constraint also made the sessions with the robot short: once the route had been conveyed by the user, the robot responded "I understand" and the experiment ended.

Another area where simulation has been employed is within the area of social robotics aimed at investigating psychological aspects of robotics. For instance to address negative emotion towards robots Nomura et al (Nomura, Kanda, Suzuki, & Kato, 2004) let users rate their experience of interacting where a simulated robot was used as a stimuli. In another study Hinds et a (2004) used a simulated robot to investigate hypotheses concerning delegation of work and trust depending on robot appearance. Goetz and Kiesler (2002) investigated the willingness of users to co-operate with a robotic instructor.

1.6.2 Simulation of multi-user scenarios

A robot that roams the corridors of an office is most likely to become a shared resource, demanding a setup for simulating a system that allow several user to co-operate for a longer period of time. Since users in a test setup are not acquainted with each other, they need some given facts about what the tasks they are supposed to solve using the robot. To be able to get momentum and involve several users a group scenario, a role-play, may be used to engage users in a task that lasts for days rather than minutes (Kanto et al., 2003). In a multi-user robotics scenario, the practical time-frame is a concern since many operators need to standby to operate the robot during the trial session. A robot working in a real environment, like an office, will spend much time travelling between different locations meaning that user studies may have to last for several hours. The burst-like character of interaction with robots in public places, reported by (Thrun, Schulte, & Rosenberg, 2000) transfers also to the case of office based service robots, e.g., when tasking and sending the robot on a mission and receiving the robot as it approaches a user with a request or offering of a transport object, users divide their attention between monitoring the robots activity and doing something else (Hüttenrauch & Severinson Eklundh, 2002).

1.7 Validity of Wizard-of-Oz simulation

Role-Play

Sometimes the Wizard-of-Oz method is criticised for not providing necessary data for evaluating practical dialogue.Wizard-of-Oz methods typically fail to involve users that bring real tasks to the system (Jönsson & Dahlbäck, 2000). This should be seen in contrast to what is generally believed, namely that the Wizard-of-Oz method is an open-ended method for collection of data on user behaviour.

Allwood and Haglund (Allwood & Haglund, 1992) have noted that the wizard operator acting a scenario is involved in roles on different levels. Thus the researcher role involves acting as a system (wizard operator) and during sessions the wizard can take on different communicative roles like the sender role, the receiver role etc. (Bell, 2003) notes that in a task scenario, like a travel agency dialogue, the wizard not only acts in a system role but in the role of a travel agent. In reality the behaviour of people acting in a real situation may be quite different from what people do in a simulated scenario even if the user believes that she is interacting with a real system.

Mismatch of Wizard Performance

When simulating a multi-modal system with many different ways for the user of providing input the cognitive load of the operators increases. There is sometimes a mismatch between what needs to be simulated and what suits the cognitive and perceptual abilities of the wizards. This phenomenon has

been noted by Fitts (1951) who proposed a list describing what people do better than machines and what machines do better than people[1]. In a simulation scenario, the problem of function allocation is twofold. First of all we need to consider the characteristics of the system we are simulating and secondly we need to think about the function allocation of the system we are using for the simulation. Following Sheridan (2000) we see may see Fitts' List as a set of accepted statements making up the foundation for how to reason when designing function allocation. According to Fitts people are better than machines at detecting small changes in the environment, perceiving patterns; improvising, memorising, making judgements and reasoning inductively. Machines are better at responding quickly to signals, applying great force, storing and erasing information and reasoning deductively (Sheridan, 2000). It is therefore important that we consider the strengths and weaknesses of machines and humans when designing and planning Wizard-of-Oz setups.

Ethical considerations

The use of the Wizard-of-Oz method provides a possible ethical dilemma for the researcher, since in the classical setup, the user is led to believe that the system is a real machine. When the truth is revealed afterwards, the test-leader normally asks for the permission to use the data. There are a number of studies that use deception; in fact it seems that the use of the term Wizard-of-Oz is always connected to the practice of deceiving the user.

Fraser and Gilbert (1991) have argued against deceiving users on ethical grounds. This ethical dilemma might be solved by telling the user that the system is simulated, making the setup part of the role-play. The validity of the results from a study that is performed in this overt fashion may be questioned but this is dependent on the purpose with the study.

Dahlbäck et al (1993) argue that for some aspects of dialogue, like the type and frequency of anaphoric references, it is important that users are deceived rather than engaging in a role-play type of scenario. We should note that the work by Dahlbäck(1993) and others in the late eighties was mostly done with systems that used text input. For spoken scenarios role-playing has been considered useful in multi-user scenarios (Kanto et al., 2003) and elicitation of error handling strategies (Skantze, 2003). Human natural language performance is automated to a great extent and therefore we may assume that phenomena on the low-level like syntax, and vocabulary used probably are little affected by the fact that the user knows that the system is simulated or not. On other levels we might find effects on the interaction. For instance it is possible that the users interaction is affected by the assumptions about the system's language capabilities, either in terms of conversational style, e.g. speech rate, elliptic information, or tasks, e.g. requests that are out of domain or directly addressing the operator behind the system (e.g. meta requests about capabilities: *Is this right? Can I do it like this?*).

The standpoint that humans treat speaking computers differently than they would treat humans is not an argument for deception per se, but an argument that promotes the use of high-fidelity simulation. However, using the Wizard-of-Oz method in the classical, deceptive manner, ensures the illusion of speaking with a computer. In a non-deceptive Wizard-of-Oz study, this illusion might be broken, something which may have an adverse effect on the result. However, it seems strange to use made up scenarios and role-play while maintaining the standpoint that we should deceive the user to the extent that the system is in fact real. In our view is that it is still an open question if deception is really necessary.

[1] The so-called MABA-MABA list.

2 Simulating the Home Tour using Wizard-of-Oz

The purpose of this study was to evaluate the dialogue developed in the initial stage of the project allowing results that are grounded in realistic scenarios to have an impact on the overall system design.

We have chosen to work with *Scenario descriptions* in order to bring an element of hypothesis testing into the process. We are therefore making a description of the scenario on an abstract level that relates a set of hypotheses with the questions we initially aim to address within the scope of the study. Scenario descriptions bring several dimensions into the overall design process. A document of this kind may contain descriptions of several tasks at different levels of abstraction ranging from an overall description of the robot's behaviour to detailed descriptions of single tasks like "specifying a new location" or "showing an object".

Analysis	Design	Implementation	Evaluation
User needs ↕ Technical assessment	→ Actions & tasks	→ Simulation tools	→ Simulation study (WOZ)
Initial Dialogue design Conforming dialogue to scenario (habituation process) Scenario descriptions – user instructions – behaviour hypotheses – robot behaviour		Dialogue tool Navigation tool Setup of logs & data collection	Pilot study (3-4 persons) Study (22 subjects) Analysis according to scenario description Explorative analysis of data

Table 1: The initial stages of the design process corresponding to the "first spiral" of the process depicted in Figure 1 (p. 7).

This information not only constrains the situation of use but also guides the analysis of the result by setting the focus on what questions we want to address in the specific scenario. On the abstract level we have chosen to describe three kinds of information for each task or phenomenon we wish to study in the simulated scenario:

User instructions providing information that enables the user to perform the intended task. This is also an important artifact, and is the primary way of influencing the user before the trial session. The instruction is either communicated verbally to the user or handed out in written form. Developing the user instruction should be an ongoing process through the pilot studies. These instructions answer the question *"What should the user do?"*.

Behaviour hypotheses regarding the designer's expectations about what the users will do in the scenario. The hypotheses answer the question *"What will the user do?"*

Robot behaviour specifies what the robot does, either as controlled by the wizard operator or as an autonomous system. The behaviour description answers the question *''What should the robot do?"*

The relation between these types of information can be described almost like a conceptual system to perform the scenario in a somewhat controlled manner. Thus the the behaviour hypotheses provides a record of what the designers expected the users to do during the scenario. The user instructions provide the means of influencing the user before and during the trial sessions that are not directly related to the robot behaviour. To achieve a controlled set up neither the instructions, nor the robot behaviour should be changed during the experiment.

2.1 Actions accommodated in the dialogue design

In the overall project description of the COGNIRON project the Home Tour is described in the following way:

> "Key-Experiment 1 is the 'Robot Home Tour'. In this experiment, a robot discovers a home-like environment and builds up an understanding of it and of artifacts in it as taught by humans. This process is open-ended, i.e., it has no completion: the robot continues to learn as it faces new situations. *A human shows and names specific locations, objects and artifacts, to the robot. The robot can engage in a dialogue in case of missing or ambiguous information.* This scenario will enable to demonstrate the capacity of dialogue, of continuous learning of space and objects. It also illustrates the possibility for the robot to take initiatives for completing its knowledge (active perception, manipulation of objects to build a sensori-motor representation)."[2].

This scenario can be characterised as kind of co-operative service discovery and configuration, stressing the way the user and robot is intended to engage in a joint effort to inform each other of relevant knowledge about the environment. This means that the user is able to discover what the robot can do and to configure it by actively providing information about:

(i) the *artifacts* present in the environment (e.g. objects and locations) and,

(ii) the *actions* that the robot can perform related to these artifacts.

In the specific instance of this scenario we are investigating the user is to guide the robot in an environment containing recognisable objects. The main task for the user is to introduce herself to the robot and to show it objects and locations. To give the user a sense of closure we have also added a validation task. This means that we are giving the user a possibility to try the functions the robot is supposed to have learned. It is also possible to end the interaction with the robot by using the conventional means of closing an interaction, (e.g. saying "Good bye"). If we consider the scenario on the abstract level there are four types of activities:

- *Introduction*. The user is able introduce herself to the robot. Directly after the introduction the robot will state that the user has been recognised and remembered.
- *Demonstration of objects and locations*. The user shows and names objects and places to the robot using speech and deictic gestures.

[2]Excerpt from COGNIRON, Annex 1: " Description of Work". Italicised by the author

- *Activation of the following behaviour*. Using the following behaviour is the intended way of controlling the movements of the robot. The follow behaviour of the robot is used to position the robot in the experiment area so that a demonstration may be performed by the user.
- *Validation of the learning process*. The user is able to find out if the robot has learned objects and locations by requesting the robot to go to locations and to find objects in the environment.
- *Closing the interaction*. The user may close the interaction by using a spoken command, for instance a parting phrase like ”good bye”.

The introduction starts the interaction and we foresee that this is a short phase leading into the main scenario. The demonstration and validation tasks can be interleaved, i.e. the user can name and demonstrate and object and immediately command the robot to find it in the environment.

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

2.2 Adapting the dialogue design

In order to be able to simulate a system that works as a unity we have used the dialogue patterns developed by our project co-partner University of Bielefeld as a starting point. We received a detailed specification of what was intended to be implemented for the system in Work Package 1.1, i.e, explicit dialogue patterns for the accommodated functions:

– Greeting,
– Closing[3],
– Person following and,
– Object and gesture detection[4].

Using these dialogue patterns we have created a dialogue design in what can be characterised as an adaptation process with the twofold goal of:

(i) Creating a system that is perceived as complete, or habitable, by the user, i.e., where there are no, or few, points where breakdowns occur because there is no “model” for handling a specific phenomena.

(ii) Familiarise the wizards to the dialogue model and the actions the system is supposed to accommodate.

The first goal is not intended to mean that the user should be able to utter an arbitrary command to the system and expect the robot to understand it, but rather that there is a relevant help message for those cases when there is no obvious response for the particular error event, according to the task model, making it possible to recover from the situation at hand. This means that sometimes the wizard must make the robot appear almost uncooperative, to stay within the boundaries of the simulated dialogue. In the following dialogue from one of the sessions the wizard keeps rejecting the users evaluative

feedback. At the final stage of the dialogue the user resorts to using the follow command, i.e., a command known to have worked before for this user.

U:	**Utterance**	**Communicative act**
U:	good work robot	FB-EVAL
R:	What did you say	REJECT
U:	good work	FB-EVAL
R:	I did not understand please rephrase	REJECT
U:	you have accomplished the task	FB-EVAL
R:	What did you say	REJECT
U:	follow me	FOLLOW

The general manner in which we have reworked the dialogue patterns results in a task-oriented dialogue model together with one or more patterns for error handling. The task-oriented model handles the behaviour that the system displays under normal (positive) circumstances. Errors that may be detected by the system (through the wizard) are to be associated with the error handling dialogue for a particular task. The relation between the different dialog patterns is that transitions between them are possible, e.g. switching from the following behaviour of the robot to dialogue concerning showing an object. This can, for instance, be done by saying “this is the cup” (see Figure 2).

2.2.1 Assuming omni-directional hearing

At a quite early stage during the implementation of the simulation system we decided not to simulate the Person Attention System as it is described by Haasch et al (2004). One of the most immediate consequences of using a mechanism like the PAS on the robot is that the robot is only able to listen to people that it is able to see, i.e., the robot needs to detect a user and explicitly turn on its recognition system before it can perceive any spoken commands. Instead we assumed that the robot could detect audio attention cues and audio commands from the user in any direction. This decision was mainly dependent on two factors:

- We believed that the users would not accept a robot that could only hear when it saw the user
- We judged it would be hard for the wizards to maintain the artificial constraint imposed by a limited PAS.

For gestures we kept with the original constraint that the users should be positioned in front of the camera.

2.2.2 Adding validation to provide a sense of closure

One of our first considerations was that the tasks greet, follow and show would make the system somewhat limited. Even if the user would be able to introduce herself to the robot and to use the follow behaviour together with the dialogue for showing objects and locations, we felt that that there was a lack of closure with respect to the robot as a agent with the ability of providing some kind of service. Hence we decided to add a validation task to make the system “complete”. The set of prompts used during the trial sessions can be found in the Appendix (p. 38).

2.2.3 Camera behaviour

Another design decision that was taken during the setup of the trial was to bring camera movements into the robot design. Initially we explored the possibilities to capture camera images. At this point we viewed the camera as a data source among others, but we quite soon realized that the movements of the camera also would affect the way the robot was perceived by the user. Hence we decided to assume a model that incorporated a moving camera (for capturing) and camera gestures with a communicative intent. The camera was controlled manually by the Navigator wizard during the follow behaviour and the search behaviour used during the show and find task. Also, when the Communicator wizard issued a spoken prompt, it was accompanied by a camera gesture displaying a gesture that makes the robot "look up" slightly towards the user (i.e., setting camera tilt to 30 degrees).

2.2.4 The follow behaviour

In general the follow behaviour worked as in the description. The movement patter of the robot behaved like a rubber band due to the reaction time of the wizard and the attempt of assuming the minimum distance of one meter. Mostly the wizard tried to follow the user by turning directly towards the users and then move in the same direction as the user. At some points the follow-wizard departed from this main pattern and instead used a deliberative model of placing the robot:

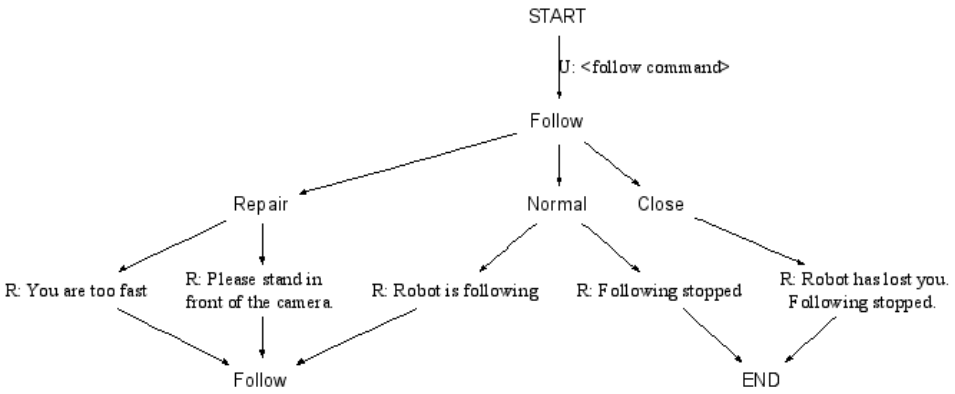
(i) Assuming a position that would have a desired effect on positioning of the user, e.g., when the robot is placed so that the user will not stand in a position that will make it possible to reveal what is going on in the wizard booth.

(ii) Positioning that will facilitate object recognition based on the condition that the wizard can anticipate what object that is about to be defined.

In figure 2 the wizard dialogue model for the Follow-behaviour is depicted. The wizard assumes that that the robot is in the follow state when the robot either has received a new follow-command (e.g. "follow me") or that the robot has recovered from an error state. In the figure the octagon labelled "FOLLOW" refers to the same state. The error states and the corresponding robot responses have been derived from the dialogue patterns. Typically an error leads to a repair, but if the user is lost the following is stopped, i.e., the wizards judge that the tracking algorithm for a real follow component would fail to re-acquire the user.

2.2.5 Adapting the "Showing" dialogue for use without a touch screen

The initial dialogue pattern for showing objects assumed a model where the robot had a touch screen. This dialogue was re-designed to handle a camera based object recognition model. The original pattern contains different prompts that were specific to the touch screen (e.g., *"please click on the ⟨OBJECT⟩ on my touch screen"*) and to the way the object recognition system would work (e.g., suggestions for recovery like *"please turn on the light"*). The model we used in the session was slightly simplified compared to the original dialogue pattern, focusing on providing smooth interaction rather than simulation of errors provided by a specific object recognition system. As it turned out we used a small set of prompt patterns in the sessions, assuming that the object recognition system would "work" unless there was some obvious error state, like several objects visible or no object visible in



Type	**Activity**	**Contribution**	**Robot Action**
Normal	Follow	**U:** Follow me!	Follow
Normal	Stop	**U:** Stop robot!	End
Normal	New task	**U:** This a cup!	End/New task
Normal	Follow	**R:** Robot is following!	Follow
Error state	**Cause**	**Contribution**	**Robot Action**
Recover	Speed ok	**R:** Robot is following!	Follow
Issue-repair	User to fast	**R:** You are too fast!	Follow
Issue-repair	User not detected	**R:** Please stand in front of the camera!	Follow
Close	User lost	**R:** Following stopped!	End

U:	Follow	Normal	Follow
R:	Robot is following ⟨*robot follows*⟩	Normal	Follow
R:	You are too fast.	Error	Repair
R:	Please stand in front of the camera	Error	Repair
R:	Robot is following ⟨*robot follows*⟩	Recover	Follow
U:	this is a book	Normal	End/New task

Figure 2: The Follow-behaviour, possible wizard actions and a and a possible dialogue scenario.

the camera view. The prompts[5] that were used for showing an object during the sessions, are listed below:

Found one ⟨OBJECT⟩
I do not know that object
Found several objects
Rearrange the objects please!

2.3 Wizard task allocation

While designing a system with the aim of simulating a real system one may consider Fitts' List (Fitts, 1951; Sheridan, 2000) when reasoning about function allocation between the wizard and the real parts of the system. It is fruitful to think in terms of what-ifs: what if the system would perform this task – what are the implications in terms of machine behaviour, and; is it possible for the operator to simulate this behaviour in a realistic manner?

In some cases human operators cannot beat the reaction time, memory capability and effectiveness of software components. In other cases human operators predict user behaviour and perform task planning in a way that far better than the system one aim to simulate. It is therefore important that we identify and address the possible mismatch between wizard performance and the performance of the simulated components. Thus, wizards need an instruction, a system to relate to when issuing commands (e.g. they need to know the boundaries for the simulated system). Furthermore wizards need to train to act as a coherent and consistent system, and finally wizard interfaces need to respond quickly so that wizards may time their responses in way that is attuned with the situation at hand.

One assumption for the whole task of controlling the robot's movements, its camera and its speech capability was that this was a job for more than one wizard. After some consideration we decided that the one wizard should control the movements of the robot platform and the other should control the dialogue behaviour. After a technical assessment of the platform capabilities we also added control of the on-board camera to the wizard tasks. The division of the wizard role is not only made on the basis of technical considerations, but also reflects the conceptual difference between moving and communicating. Hence one wizard role is that of the "Navigator" and the other is the "Communicator". In the setup the Navigator wizard also acts as test leader towards the user. The Communicator wizard acted as "Technician" towards the user. The test leader and technician roles could be switched. In figure 3 an overview is shown.

Defining the wizard task as the two subtasks: navigation and communication has been the preferred model of others that have attempted at simulating multi-modal human-robot interaction. For instance, Perzanowki et al (2003) used a this division of labour in a pilot study aimed at collecting multi-modal data. One wizard was controlling the navigation of the robot; the other acted as the robot's speech interface using a headset microphone attached to a sound modulator to produce a robotic voice. Two wizards were also employed in the study by Green et al (2000). During the sessions one wizard was responsible for the physical movement of the robot and the other provided dialogue capabilities by playing prompts using the on-board speech synthesiser.

To provide multi-modal dialogue capabilities for the robot the Communicator wizard has a tool (figure 4) that provides output from a large set of phrases. Since the dialogue interface simulated here is aimed at a task-oriented type of dialogue we have assumed that phrases may have two functional

[5]The ⟨OBJECT⟩ placeholder was expanded to a set of phrases containing the known objects in the dialogue tool

Wizard function	Wizard 1	Wizard 2
Visible function	Role: Test leader – Welcoming user	Role: "Technician" – answers questions related to speech system
Control function	Role: Navigator – Executes navigation according to intended follow behaviour – Manoeuvring the on-board camera	Role: Communicator – Dialogue management

Figure 3: The wizards' visible and covert roles

types: task-related and general feedback. This is reflected in the interface where the left table holds task-related phrases and the right table holds feedback phrases.

To handle phrases containing locations or objects we have added columns to hold objects and locations. When a task-oriented phrase containing a type marker (e.g."*LOCATION*") is selected a dialogue window containing the set of expanded phrases for the possible location is displayed. This makes it possible to produce hundreds of phrases with a few mouse clicks.

The wizards also have access to a list showing objects that have been mentioned during the session. The list was added to the interface after the pilot sessions. A stop watch timer was also added to keep track of the length of the sessions. The tool also contains fields for commands to produce simultaneous robot actions, e.g., sending a move command while letting the robot say "moving forward". This feature was used to provide some camera movements corresponding to communicative feedback (e.g. looking slightly upwards when saying "Hello").

Using the navigator tool the wizard is able to directly control the robot's movements using a standard type gaming joystick. For this task the most important feedback to the wizard is provided by directly monitoring the robot itself. The on-board camera image also gives some information that can be used to decide where the robot is looking. The gaming joystick (a Wireless Logitech Wingman) provides a set of programmable buttons that can be used to implement shortcuts that facilitate quick responses and speed up wizard reactions.

2.3.1 The experimental environment

The wizard scenario was set up in an experiment environment called the "Living room" located in the robot lab at CAS/KTH (see figure 5a). This is an office room (5 X 5 metres) furnished with typical Swedish furniture[6] to resemble a living room in a normal Western European house or apartment. The room was provided with a set of everyday objects. In each upper corner a network web-cam was placed to capture images overlooking the whole room. On a table in one corner of the room a screen was put up to cover the two computers that were used by the wizards. We wanted to prevent the users from directly seeing what the wizards were doing. In a more classical setup for a Wizard-of-Oz experiment the wizards are not visible. In this setup we have lifted this constraint, because the wizards

[6] From IKEA.

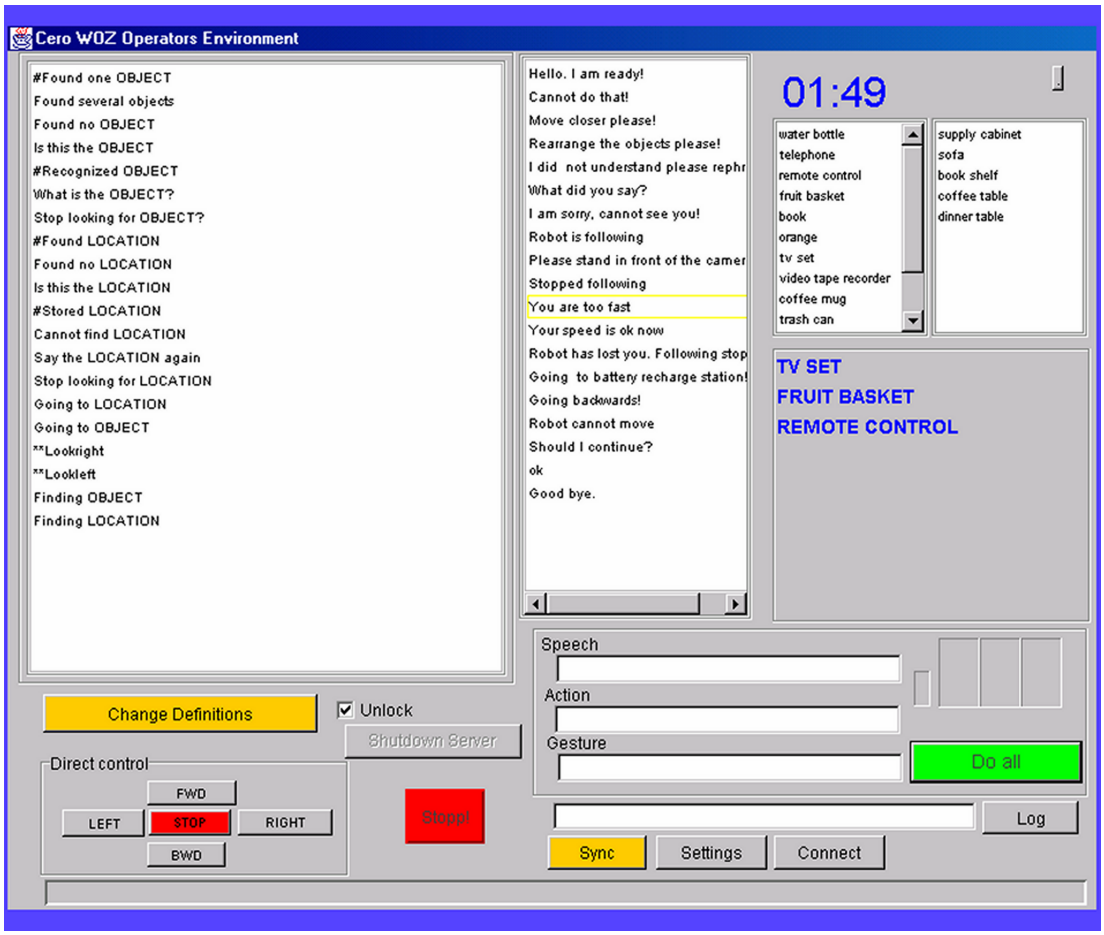


Figure 4: The Dialogue production tool.

need access to a significant amount of information about the positioning, body posture, gesture and commands that are being used by the subject. Our experiences from earlier studies, and the pilot studies, is also that users accept the fact that an additional staff member (termed “technician”) is required to maintain the robot during the session.

2.3.2 The robot

The robot we used for the trials we used a ActiveMedia Peoplebot[7]. The robot (figure 5) was equipped with four visible microphones, two of which were attached to metal wires on the top of robot to collect less noise from the robot platform. Two other microphones were less prominent but yet visible to the subjects. The robot also had a pan-tilt video camera providing a simple gaze mechanism for the robot. The gripper was concealed by a digital sound recorder used for the collection of on-board sound. One prominent feature of the lower part of the robot was a SICK laser range finder with the (standard) clear blue colour. On the upper and lower part of the robot two sets of sonars sensors were attached. The sonars were switched off during the experiment to reduce the noise from the robot. Still the level of noise from the fan of the robot’s on-board computer and the motors was considerable. Given the

[7] www.activemedia.com

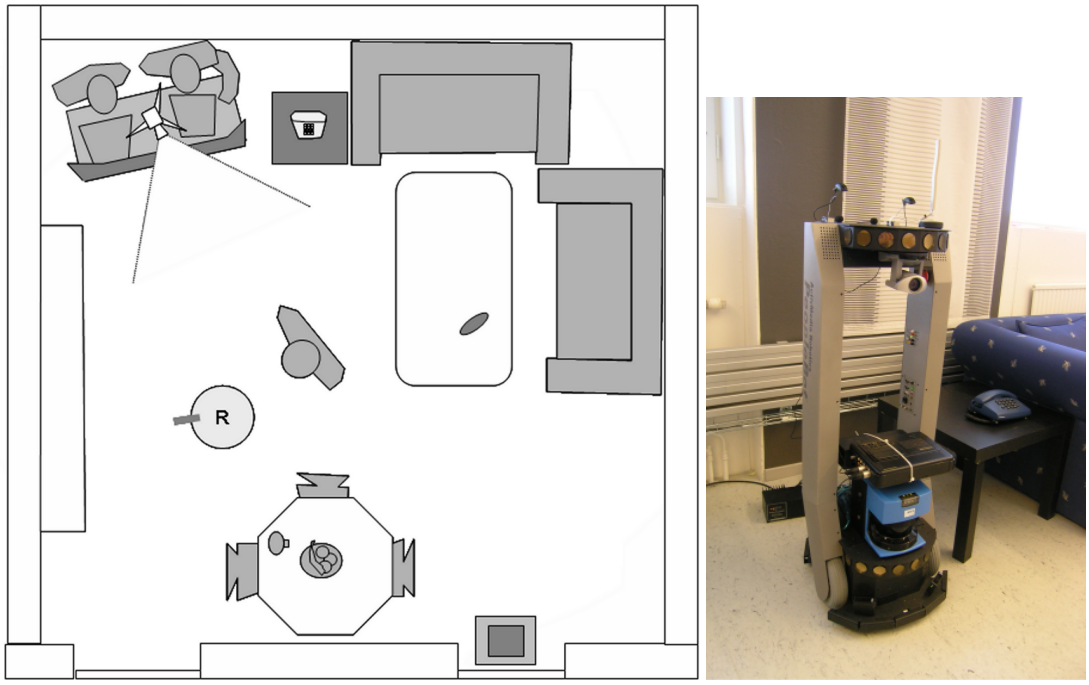


Figure 5: a) The wizard set-up in the "Living room" environment and the b) robot used in the trial sessions.

time frame and technical risk involved we decided that we would not attempt to reduce the sound by modifying the cooling system of the robot.

2.3.3 Participants and test procedure

Initially we performed a formative pilot study with a few staff members in order to fine tune the setup. In the next phase we recruited 22 test persons among students on the KTH campus. This means that there is a bias towards well-educated young people in the study, but since the aim of the study is primarily explorative we have accepted this circumstance.

Upon arrival the subject was greeted by the test leader and offered a cup of coffee. Then the test leader informed the subject of the purpose of the study, without revealing that the wizards were controlling the system. Instead the wizards were described as "technicians" with the purpose of controlling the technical setup and making "online annotations". During the trial there were three researchers present; one acting as test leader/navigator; one acting as communicator; and one acting as observer. During the setup the observer was positioned in one of the sofas taking notes.

After the introduction the subject signed an agreement giving consent to storing of personal information[8]. The subject was then instructed to read the written user instruction (see the Appendix). After the subject had finished reading the instruction the test leader addressed any questions or requests. Then the test leader[9] gave the following demonstration standing in front of the robot:

[8] Required according to Swedish law.
[9] The navigation behaviour of the robot was controlled by the communicator wizard.

TL: Hello robot!
R: Hello I am ready
TL: Follow me
⟨*robot follows TL*⟩
⟨*TL stands in front of book shelf*⟩
⟨*TL points at book*⟩
TL: This is a book
R: Found one book
TL: Go to the battery re-charge station
TL=test leader, R=robot

When the robot made its way back to the battery station the test leader asked the user if there were any further questions and got in position position behind the screen. Then the user was given the cue that it was ok to start the interaction. About this time the on-board sound recordings, the acquisition of laser data and the video camera were started. Then the test went on for approximately fifteen minutes.

During the interactions the users sometimes addressed the test leader. Depending on the kind of question or issue the test leader answered it or asked the technician. For instance, when there was some question about the dialogue system, e.g. “What can I say?”, the test leader turned to the “technician” asking him to address the request of the user. The response given was to encourage the user to try to address the robot using her own words.

The session ended when the user, or in same cases the test leader, told the robot go to the battery station and told the robot “Good bye”. This was either at the initiative of the user or prompted by the wizards. The method for doing this was to let the robot play a sound that signalled that the batteries were running low. The test leader then prompted the user to send the robot to the battery station. The test leader said that the user was over and asked the user to fill out the questionnaire. After finishing the questionnaire the test leader and the “technician” engaged in a post session interview. This gave the subject the opportunity to suggest things or to comment on the robot interaction in an open-ended manner, still under the impression that the robot was for real.

The last part of the post session interview was the actual debriefing of the user. This was initiated by the test leader by asking the user to comment on one of the questions of the questionnaire, concerning who was controlling the robot. The reason for this was to check if the subject had any suspicions about the setup. After establishing this the test leader revealed the truth. At this point the post session turned into debriefing which normally went on for a couple of minutes.This meant that we asked the subject if there were any concerns due to the fact that we had simulated the robot’s interface. We also said that we were doing this with good intention and that we appreciated his or her involvement in our research. The user was then rewarded a cinema ticket and left.

2.4 Data Collection

In order to be able to analyse the study from different perspectives data of various types were collected. First of all video from the overall scene was recorded, using a Mini DV camera. This camera was placed on a tripod behind the wizards’ screen and handled by the Communicator wizard. This camera recorded video (Mini DV) and audio and was operated by the dialogue wizard. The camera was equipped with a wide angle lens in order to capture as much of the scene as possible and minimising

the need to pan the camera during the trial sessions. We also collected images from four network web-cams, placed in each corner of the room. The purpose of collecting this data was to get an overview of the scenario, and to figure out what went on when the main camera was occluded. The frame rate of the network cameras was approximately one frame per second, depending on network traffic and load of the servers where the images were stored.

Audio from two different sources was collected: the sound from the wizard's video camera and the sound from the stereo microphones placed on top of the robot. The on-board sound was recorded on a digital audio recorder placed on the robot gripper. The sound was stored in .WAV-format.

The microphones were facing forwards, and sat on top of the robot approximately on a hight of 125 cm and with a distance of 40 cm between them (see Fig 5). The microphones were also arranged to minimise the sound from the robot platform.

A log was kept of the commands that the server received. The format of the logs contained information about the clock time the command was executed and the synctime, i.e., the time in seconds and milliseconds from initialisation of the server. Logs that were kept on the individual clients have not been collected. The log format is shown below:

Log format:[10]

```
<clocktime= <DATE> synctime=1361.154 val=''campantilt 0 30''>
<clocktime= <DATE> synctime=1361.158 val=''saytext Hello. I am ready!''>
<clocktime= <DATE> synctime=1366.331 val=''campantilt 0 30''>
<clocktime= <DATE> synctime=1366.334 val=''saytext Robot is following''>
<clocktime= <DATE> synctime=1368.211 val=''campantilt 0 0''>
<clocktime= <DATE> synctime=1368.386 val=''drive 5.0 0.
```

We also collected data from the SICK laser range finder attached to the robot. The range finder is placed approximately 30 centimetres over the floor and covered an area of 180° in front of the robot.

Data type	**Description**
Video	Mini DV video (PAL) from Wizard's point of view.
Four network cameras	Still images with $\sim$ 1 fps from four different angles. Images are stored in JPG-format.
Camera audio	Audio from the Wizard's camera video tape (48 KHz)
On-board audio	Audio from the on-board (Wav-format, Stereo 16 KHz, 16 bit uncompressed).
Server log	Log entries of commands received from the wizard tools.
Laser	Data from the range finder facing forward on the robot

2.5 Annotation procedure

Since we have collected many hours of interaction we need to carefully consider a level of annotation that gives enough input for further analysis. As a starting point we have therefore annotated the first five sessions in a number of ways to get a clear picture of the amount of work required to annotate the rest (15 sessions) and what analyses that are possible to do, given the level of annotation.

The master file determining the synchronisation of audio and video is the video recording from the DV cam-corder. After capturing the video using Adobe Premiere 1.5 we edited the on-board audio

[10] ⟨DATE⟩ are entries of the format `Wed Sep 08 12:28:27 CEST 2004`

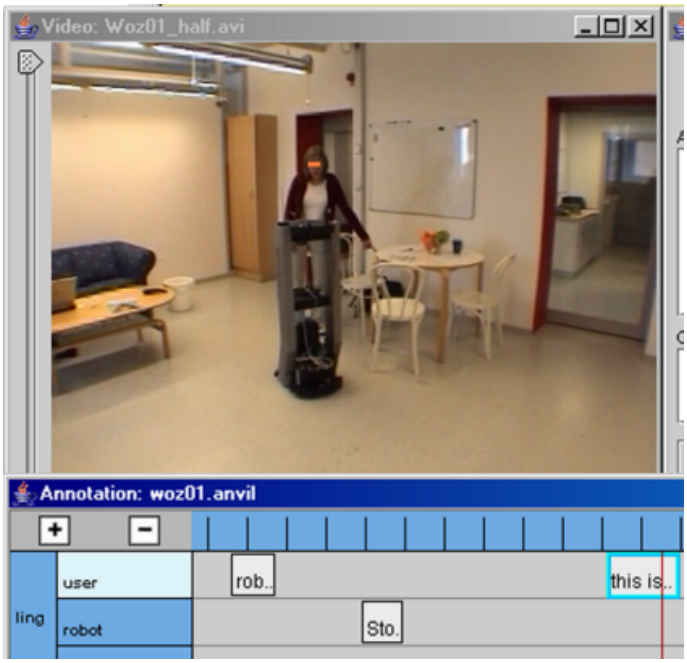


Figure 6: The Anvil view

files so that they were synchronised, and of equal length. In some cases that meant that we had to add a blank movie sequence to the DV-file, but in most cases the audio files were cut at the start and end points.

The user sessions are annotated on the utterance level, defining an utterance as something which to the transcriber seems to be a coherent sequence of speech. This means that sometimes the word “robot” followed by a *significant* pause and a command is treated as two separate utterances (e.g., “robot” and “follow me”) . In other cases the similar constructions, starting with “robot” is taken to be one utterance, e.g., “robot follow me”.

During the user sessions an automatic log was kept whenever the wizard made the robot speak. By inspecting a spectrogram view of the on-board audio file the offset between the time of occurrence for the first robot utterance in the recorded session and the log time of the corresponding command in the log was established. Using the offset value a simple script was written (in Perl) to generate a file of transcription labels (.lab) to be used in the Wavesurfer[11] tool. In the spectrogram view of Wavesurfer each start and end point for the labels of the utterances was edited based on their tentative position calculated from the log entry. This was done by dragging start and end markers in the graphical user interface. This practice reduces the amount of work needed to adjust the synchronisation of the utterance and the label, compared to typing in values. During this stage in the process we also manually inserted annotations of the user’s utterances. The result of this process stage was then used to generate an Anvil transcription file according to a specific XML-schema (see the Appendix).

[11] http://www.speech.kth.se/wavesurfer

Type	Description	Verbal	Task	Occurrances
Request Attention	User requests attention from robot by using voice (e.g., hello), noise (e.g., clap, whistle) and or gesture (wave)	Attention	Any task	W03: 81, W03: 144, W03: 152 W05: 9, 469, 471
Deixis: Focusing	User focuses her/his posture to enhances and increase the effect of pointing	Assert	Show	W01:269, W02:395, W03: 316
Deixis: Point and hold	User points at a specific point for a long time. User also touches object, to save energy.	Assert	Show	W01: 29-37, 76-92, 125-139, 172-187, 266-280, 305-335, 372-284 W02: 125-130, 168-178 W04: 28-34, 36-45,119, 244-259, 309-317
Deixis: Point to robot	User points to robot when uttering "stop". The gesture is executed over a short time $\sim$1 s.	Stop	Follow	W01: 21, 117, 263, 302, 370
Deixis: Hold object	User holds object while specifying object.	Assert	Show	W03: 180. W04:140-148
Configuring+Deixis	User rearranges the object at a location. Then points to the object.	Assert	Show	W02: 273-301, W04:150
Probing: Repeat phrases	User repeat utterance. Varying phrasing.	Assert	Show	W03: 47 57, W03: 66-70, W03:317-322 W05: 263-291

The first and second columns (*Type*, *Description*) give the type and description of the patterns. The third column (*Verbal*) states the speech act that can be associated with the patterns. The fourth column (*Task*) describes the task associated with the pattern. The fifth column (*Occurrances*) contains the references to tokens of the pattern within the recorded material. References are on the form: *W#:s-s* denoting the wizard session number, start and end time in seconds from the beginning of the session recording.

Table 2: Some patterns occurring in the first five trial sessions.

3 Preliminary findings

The analysis of the recorded trial sessions has only begun. We have gone through the first five videos several times, noting interesting patterns. In the following we will present examples that illustrate some of these patterns (see Table 2).

Request attention: In the first five sessions we have noted different ways the user may request the attention of the robot:

- Verbally (e.g. "Hello robot")
- Gesture (e.g. wave)

- Sound (e.g. “Whistle”)

The greeting phase is the most common way to request attention. In the first five sessions the users greeted the robot and got the response that the robot was ready (“Hello I am ready”).

Session 1:	**U:**	robot
	R:	Hello I am ready.
Session 2:	**U:**	hey
	R:	Hello I am ready
Session 3:	**U:**	hello robot
	R:	Hello. I am ready
Session 4:	**U:**	hello robot
	R:	Hello. I am ready
Session 5:	**U:**	hello robert*
	R:	Hello. I am ready

*The user used the proper noun ’Robert’

Requesting attention by clapping hands and whistling was only seen in one of the sessions. Waving in front of the robot or getting in front of the camera, perhaps to establish eye-contact, was somewhat more frequent.

U:	follow me	FOLLOW
R:	Robot is following	FOLLOW
U:	CLAPS HANDS	REQ-ATT
U:	WHISTLES	REQ-ATT
R:	Robot is following	ACK
U:	this is a chair	ASSERT
	robot moves too close to user (error in follow behavior)	

Deixis – Point and hold: During pointing sequences some users seemed to adopt a strategy that can be termed “point-and-hold” starting with a pointing gesture by the user that was hold in an almost static pose until the robot gave some feedback. The hold-phase of such an event could last for several seconds. In some cases we observed poses that seemed to be uncomfortable such as stooping or bending the upper body while maintaining the pointing gesture. It seems that touching the object then worked as a way of saving energy during the hold-phase of a pointing sequence.

Deixis – focusing: Another patterns that is interesting was how the users sometimes seemed to focus their pointing gesture. A typical focusing consists of a pointing gesture, then when there is no response from the robot, or the camera points in another direction, the user moves the hand closer to the object. This is seen in Figure 7 and in Figure 8. The first example the user focuses the gesture just before verbally indicating the object. In the second example, from one of the pilot sessions, (Figure 8 p. 30) the user notices that the camera is looking in an other direction and takes a quick step to the left. Then she moves closer to the object while maintaining the pointing gesture.

Deixis – holding objects: Small objects, such as the flash light and the bottle of glue were sometimes moved around before being shown to the robot. Since these objects were held by the users during the configuration phase the user did not release them as they were instructed. This was considered to be an error and an a repair was issued by the communicator wizard. The prompt “please rearrange the objects” was used in these cases and typically the user then let of of the object and issued a pointing gesture instead.

Image 1-3, numbered left to right.

U_1: robot stop (Image 1)
R_2: Stopped following
U_3: this is a telephone (Image 3)
R_4: Found one telephone

Figure 7: The user points at the robot while saying ”stop” to it (Image 1, U_1). Then the user points at the telephone (Image 2). Moving the pointing hand slightly closer before uttering “this is a telephone” (Image 3, U_3).

Probing: Some users were attempting to find the boundaries of the robot’s capability. Probing the boundaries seems to be triggered by:

- The robot explicitly asks the user to rephrase
- The discovery of a new function, e.g. assuming that the robot is capable of responding to commands for directive navigation (i.e., left, right, forwards etc).
- After the first instance of placing an object, the fact that the robot had a search behavior seemed to trigger more creative use of the objects present, e.g. placing objects on non-standard places.

In the dialogue model we were simulating we only responded affirmatively to one type of directive navigation: going backward. The reason was that we thereby made it possible to recover from situations where the robot got stuck in some way. If other directive commands were issued by the user we responded in a negative way. In the following example a breakdown has occurred, leaving the robot (too) close to the user. The user commands the robot to go back. When the robot complies the user consequently infers that other directive commands should be possible and addresses the robot with several varied commands (in boldface below).

Image 1-6, numbered left to right (in a 2 X 6 grid).

U: This is an orange (Image 2)
R: Found object (Image 6)

Figure 8: In this sequence (from the pilot study) the user first points to the orange (1-2). When the camera looks in another direction, the user takes a quick step to the left (4). Then she moves closer, still pointing (5-6). The image of the camera has been enlarged in the left corner of each image. A black rectangle has been drawn in the picture between the user's body and the object. It is intended to give an idea of the distance between the user and the object.

U:	back	**DIRECTIVE**
U:	faster	**DIRECTIVE**
R:	ok	ACK
U:	stop	STOP
R:	ok	ACK
U:	forward	**DIRECTIVE**
U:	very fast	**DIRECTIVE**
R:	Cannot do that	REJECT
U:	forward	**DIRECTIVE**
U:	fast	**DIRECTIVE**
R:	Cannot do that	REJECT
U:	fast forward	**DIRECTIVE**
R:	Cannot do that	REJECT
U:	forward	**DIRECTIVE**”
R:	Cannot do that	REJECT

Probing – re-configuring the environment: Many users moved objects around when attempting to probe the robot's ability to find an object. Here the creativity was large.

Sometimes the users also display their intentions before actually stating anything directly to the robot. One example of this is shown in Figure 9. The spoken interaction during this interaction is rather sparse but from the activity displayed by the user it is evident that a case of showing the robot an object is going to take place. In Figure 9 the user is rearranging the the chairs present in the room during the follow behavior. It is hard to say if the user has actually decided what the next task is. A careful guess is that the user has not actually decided on the next move before moving out on the floor. However, the user clearly touches both the chairs, as if making a choice between them, before heading back to fetch one of them. This chair is then positioned on the floor, in front of the robot, with the seat facing towards it. The seat of the chair seems to be used in a conventional way, as if the user assumed that the robot would take the facing of the chair into consideration.

Image 1-12, numbered left to right (in a 3 X 4 grid).

Figure 9: *User grasps a chair (2), then releases the chair. Then touches the next chair while moving out on the open floor. Beside the table the user keeps close to the table (2-3), and moves one of the chairs (4-5). The robot follows the user by slowly turning (2-5). The user turns back to see what the robot is doing (6). The user reaches for the last chair (6-7), positioning it in the middle of the floor so that the seat faces the robot (9-11). Finally the user points at the chair (12).*

4 Conclusions and future work

This report aimed to describe the way we have worked with hi-fi simulations within the initial phase of the project. It also describes the data collection and some very preliminary results from the analysis of the data.

In the next phase of the project we aim to develop methods and tools for performing theoretically informed analysis of data from user studies. The resulting methods and tools will combine the analysis of dialogue and nonverbal/spatial aspects of interaction. This work will yield, among other things, an annotated video corpus of robot directed speech and gestures which we aim to make available for internal use in the project.

With these data, different dialogue phenomena central to cooperative service discovery and configuration can be studied, in particular:

- Miscommunication and breakdowns
- Generic user behaviour/interaction styles

Analysis of user data will allow us to enhance and reconsider the criteria for communicative success developed within other related areas like unimodal dialogue systems, speech enhanced graphical user interfaces and uni-modal gesture based interfaces. Initially we investigate phenomena to develop criteria for avoiding and handling miscommunications and breakdowns.

We will also study dialogue structures to enable human augmented mapping and basic dialogues about spatial relations that will fit into the scenario of mapping. To begin with, this work will proceed by generating synthetic dialogues and using partial simulations to study dialogue strategies. A first test will thus be to perform a new user study of the home tour scenario in which the Wizard-of-Oz functionality is replaced for interaction using a set of robot behaviour.

References

Allwood, J., & Haglund, B. (1992). *Communicative Activity Analysis of a Wizard of Oz Experiment* (Tech. Rep.). Department of Linguistics, Göteborg University.

Antoniol, G., Cattoni, R., Cettolo, M., & Federico, M. (1993). Robust Speech Understanding for Robot Telecontrol. In *Proceedings of the 6th International Conference on Advanced robotics* (p. 205-209). Tokyo, Japan.

Bell, L. (2003). *Linguistic adaptations in spoken human-computer dialogues: Empirical studies of user behavior*. PhD Thesis, KTH Royal Institute of Technology. (TRITA-TMH 2003:11)

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, *23*(1), 13-31.

Casper, J., & Murphy, R. R. (2003). Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics Part B*, *33*(3), 367-385.

Cheepen, C., Gilbert, N., Failenschmid, K., & Williams, D. (2002). Guidelines for the design of advanced voice dialogue. (Department of Sociology, University of Surrey, UK and Vocalis Ltd)

Corazza, A., Federico, M., Gretter, R., & Lazzari, G. (1993). Design and acquisition of a task-oriented spontaneous-speech data base. In V. Roberto (Ed.), *Intelligent perceptual systems* (p. 196-210). Heidelberg, Germany: Springer Verlag.

Crandall, J. W., & Goodrich, M. A. (2003, September 16-18). *Measuring the intelligence of a robot and its interface.*

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

Ehn, P., & Kyng, M. (1992). Cardboard computers: mocking-it-up or hands-on the future. In (pp. 169–196). Lawrence Erlbaum Associates, Inc.

Fitts, P. M. (1951). *Human engineering for an effective air navigation and traffic control system*. Washington, DC: National Research Council Committee on Aviation Psychology.

Fraser, N. M., & Gilbert, G. N. (1991). Simulating speech systems. *Computer Speech & Language*, *5*(1), 81-99.

Fry, J., Asoh, H., & Matsui, T. (1998). Natural Dialogue with the JIJO-2 Office Robot. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, *2*, 1278-1283.

Goetz, J., & Kiesler, S. (2002). Cooperation with a robotic assistant. In *CHI'02: CHI'02 extended abstracts on Human factors in computing systems* (pp. 578–579). ACM Press.

Green, A. (2001, September). C-Roids: Life-like Characters for Situated Natural Language User Interfaces. In *Proceedings of the 10th IEEE International Workshop on Robot and Human Interactive Communication Ro-Man01*. Bordeaux/Paris.

Green, A., Hüttenrauch, H., Norman, M., Oestreicher, L., & Severinson Eklundh, K. (2000). User-centered design for intelligent service robots. In *Proceedings of 9th IEEE international workshop on robot and human interactive communication.* Osaka, Japan.

Green, A., & Severinson Eklundh, K. (2001, September). Task-oriented Dialogue for CERO: a User-centered Approach. In *Proceedings of 10th IEEE international workshop on robot and human interactive communication.* Bordeaux/Paris.

Haasch, A., Hohenner, S., Huewel, S., Kleinehagenbrock, M., Lang, S., Toptsis, I., et al. (2004, May 21). BIRON – the Bielefeld Robot Companion. In *Proceedings of ASER 2004 - 2nd International Workshop on Advances in Service Robots.* Stuttgart, Germany.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, 151–181.

Hulstijn, J. (2000). *Dialogue models for inquiry and transaction*. PhD Thesis, University of Twente.

Hüttenrauch, H., & Severinson Eklundh, K. (2002). Fetch-and-carry with cero: Observations from a long-term user study with a service robot. In *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication Ro-Man2002* (p. 158-163). Berlin, Germany. (September 25-27)

James, F., Rayner, M., & Hockey, B. A. (2000). *"Do that again": Evaluating spoken dialogue interfaces* (Tech. Rep.). RIACS Technical Report #00.06.

Jönsson, A., & Dahlbäck, N. (2000). Distilling dialogues - a method using natural dialogue corpora for dialogue systems development. In *Proceedings of 6th applied natural language processing conference* (pp. 44–51).

Kanto, K., Cheadle, M., Gambäck, B., Hansen, P., Kristiina Jokinen, H. K., & Rissanen, J. (2003). Multi-session group scenarios for speech interface design. In C. Stephanidis & J. Jacko (Eds.), *Human-Computer Interaction: Theory and Practice (Part II)* (Vol. 2, p. 676-680). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Karsenty, L. (2001). Adapting verbal protocol methds to investigate speech systems use. *Applied Ergonomics*, *32*, 15–22.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, *2*(1), 26–41.

Lauria, S., Bugmann, G., Kyriacou, T., & Klein, E. (2002). Mobile robot programming using natural language. *Robotics and Autonomous Systems*, *38*(3-4), 171-181.

Lövgren, J. (2004, November+December). Animated use sketches as design representations. *interactions*, *6*(22-27).

Malhotra, A. (1975). *Design criteria for a knowledge-based english language system for management: an experimental analysis.* (Tech. Rep. No. MAC TR-146). MIT.

Maulsby, D., Greenberg, S., & Mander, R. (1993, April). Prototyping an Intelligent Agent through Wizard of Oz. In *INTERCHI'93* (pp. 277 – 282).

Mival, O., Cringean, S., & Benyon, D. (2004). Personification Technologies: Developing Artificial Companions for Older People. In *Proceedings of the 2004 conference on Human factors and computing systems: CHI2004.* ACM Press.

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. In *Proceedings of 13th IEEE International Workshop on Robot and Human Interactive Communication Ro-Man2004*.

Oviatt, S. (1996). User-centered modeling for spoken language and multimodal interfaces. *IEEE Multimedia,, 3*(4), 26-35.

Perzanowski, D., Brock, D., Adams, W., Bugajska, M., Schultz, A. C., Trafton, J. G., et al. (2003). Finding the FOO: a pilot study for a multimodal interface. In *IEEE International Conference on Systems, Man and Cybernetics* (Vol. 4, p. 3218-3223).

Perzanowski, D., Schultz, A., Adams, W., Marsh, E., & Bugajska, M. (2001). Building a multimodal human-robot interface. *Intelligent Systems, IEEE [see also IEEE Expert]*, *16*(1), 16-21.

Reeves, L. M., Martin, J.-C., McTear, M., Raman, T., Stanney, K. M., Su, H., et al. (2004). Guidelines for multimodal user interface design. *Communications of the ACM*, *47*(1), 57-59. (SPECIAL ISSUE: Multimodal interfaces that flex, adapt, and persist)

Sakamoto, D., Kanda, T., Ono, T., Kamashima, M., Imai, M., & Ishiguro, H. (2004). Cooperative

embodied communication emerged by interactive humanoid robots. In *Proceedings of 13th IEEE International Workshop on Robot and Human Interactive Communication Ro-Man2004.*

Salber, D., & Coutaz, J. (1993). Applying the Wizard of Oz Technique to the Study of Multimodal Systems. In *EWHCI* (pp. 219–230).

Scholtz, J. (2002). Evaluation methods for human-system performance of intelligent systems. In E. Messina & A. Meystel (Eds.), *Proceedings of the 2002 Performance Metrics for Intelligent Systems (PerMIS) Workshop*.

Sheridan, T. B. (2000). Function allocation: algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies*, *52*(2), 203-216.

Skantze, G. (2003). Exploring human error handling strategies: Implications for spoken dialogue systems. In *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems* (p. 71-76).

Thrun, S., Schulte, J., & Rosenberg, C. (2000, July/August). Interaction with mobile robots in public places. *IEEE Intelligent Systems*, 7–11.

Torrance, M. C. (1994). *Natural Communication with Robots*. Unpublished master's thesis, MIT Department of Electrical Engineering and Computer Science.

Appendix

A Communicative acts

The labels used to tag communicative acts used in the examples are similar to the categories used in the DAMSL[12] speech act tag set. The list below is preliminary and is neither complete nor final.

Tag	Description	Example
REJECT	Reject the proposition by the last speaker.	*R: Cannot do that*
ACK	Acknowledge the proposition by the last speaker	*R: Ok.*
DIRECTIVE	An action directive command aimed to influence the behavior of the listener.	*U: Go backwards.*
ASSERT	A proposition asserting some property.	*U: This is a chair.*
STOP	Stop is a frequently directive command used to stop the robot's movements.	*U: Stop*
FOLLOW	Follow activates the follow-behavior of the robot.	*U: Follow me*
REQ-ATT	Any utterance or gesture aimed at grabbing the attention of the robot.	*U: Hello robot!* + Wave
FB-EVAL	An utterance providing positive or negative evaluation.	*U: Good work!*

[12] See `http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/`

B Prompts

These prompts were used in the dialogue tool. They are ordered thematically below, but in reality, as some of them were used frequently they were ordered differently in the tool. The placeholders (e.g., ⟨LOCATION⟩ and ⟨OBJECT⟩) in the phrases below were expanded with the objects and locations respectively in the dialogue tool. This is gave a compact representation of phrases, but allowed for hundreds of different phrases to be constructed with only a few steps in the graphical user interface of the dialogue tool.

Greeting
Hello. I am ready!

Following
Robot is following
Robot has lost you. Following stopped.
Stopped following
You are too fast
Your speed is ok now

Locations
Found ⟨LOCATION⟩
Stored ⟨LOCATION⟩
Cannot find ⟨LOCATION⟩
Finding ⟨LOCATION⟩
Found no ⟨LOCATION⟩
Is this the ⟨LOCATION⟩
Going to ⟨LOCATION⟩
Say the ⟨LOCATION⟩ again
Stop looking for ⟨LOCATION⟩
Should I continue?
Going to battery recharge station!

Objects
Recognized ⟨OBJECT⟩
Found one ⟨OBJECT⟩
Going to ⟨OBJECT⟩
Finding ⟨OBJECT⟩
Found no ⟨OBJECT⟩
Is this the ⟨OBJECT⟩
Stop looking for ⟨OBJECT⟩?
What is the ⟨OBJECT⟩?
I do not know that object
Found several objects
Rearrange the objects please!

Directive
Going backwards!

Closing
Good bye.

Repairs
I did not understand please rephrase!
What did you say?
Cannot do that!
Robot cannot move
I am sorry, cannot see you!
Move closer please!
Please stand in front of the camera

Objects and locations
apple
banana
chair
fruit
book
book shelf
coffee mug
coffee table
dinner table
fruit basket
ok
orange
remote control
sofa
supply cabinet
telephone
trash can
tv set
video tape recorder
water bottle

B.1 User instruction

Human Robot Interaction

The user trial you are about to take part in is intended to explore human-machine interactions with service robots. For the next 45 minutes (approx.) you will drive and evaluate a robot prototype. The experiment will be con-ducted in English, but otherwise you will need no technical experience.
We would also like to stress that we are not testing your performance, but the performance of the user interface of the robot. The trial will be recorded and used for further analyses. We would like to stress that your participation is entirely voluntary and that you at any point during the experiment may choose to quit the experiment.
Welcome!

Training a service robot

Imagine that you have bought a service robot that is to perform simple duties in your home. Maybe you have broken your leg while when you went skiing or maybe you are waiting to get you're your hip joints replaced. Anyway, the robot is here, fresh out of the box and now you have to train it to suit your environment.
The robot can understand speech instructions used when training it as well as pointing gestures. Below the features of the robot system will be explained to you together with some of its limitations.

- To interact with the robot, place yourself within 1.5 to 3 meters in front of it.
- To begin the interaction you can greet the robot. Using its camera the robot will learn to recognize your face so that will be able to follow you.
- To end interaction with the robot say "Good bye"

Following

The robot is able to follow you on your command. Position yourself so that the robot is able to detect you with its camera, and then say "Follow me".

Teaching it places

Direct the robot to places that it is to learn using the follow-function and then name the locations to the robot. Once the robot is at the right position you can define the location by naming it to the robot.

Teaching it objects

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

Locate places and objects

Once the robot has been shown an object or a location it may be told to discover it again. The objects may be placed in another place than where the robot learned about it.

- Ask the robot to transfer to a place that you know it should have learned.
- Put an object in a location and ask the robot to discover it.

Test instruction

On the next page, places and objects that you should show to the robot are depicted. To train the robot, use the functions for following, teaching and finding locations and objects.

C The Anvil XML schema

This is a preliminary version containing three groups: linguistic information, gestural information and actions.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<annotation-spec>
 <head></head>
 <body>
  <group name="ling">
    <track-spec name="user" type="primary" >
      <attribute name="contribution" valuetype="String"/>
     </track-spec>
    <track-spec name="robot" type="primary">
      <attribute name="contribution" valuetype="String" />
    </track-spec>
 <track-spec name="tl" type="primary">
      <attribute name="contribution" valuetype="String" />
    </track-spec>
  </group>
  <group name="gest">
    <track-spec name="user" type="primary" >
      <attribute name="contribution" valuetype="String"/>
     </track-spec>
    <track-spec name="robot" type="primary">
      <attribute name="contribution" valuetype="String" />
    </track-spec>
    <track-spec name="robotcamera" type="primary">
      <attribute name="contribution" valuetype="String" />
    </track-spec>
    <track-spec name="other" type="primary">
      <attribute name="contribution" valuetype="String" />
    </track-spec>
  </group>
  <group name="actions">
    <track-spec name="user" type="primary" >
      <attribute name="contribution" valuetype="String"/>
     </track-spec>
    <track-spec name="robot" type="primary">
      <attribute name="contribution" valuetype="String" />
    </track-spec>
    <track-spec name="other" type="primary">
      <attribute name="contribution" valuetype="String" />
    </track-spec>
  </group>
 </body>
</annotation-spec>
```