User Study: Talking to Computers

Markus Berg
University of Applied Sciences Wismar
Wismar, Germany
markus.berg@hs-wismar.de

Petra Gröber
University of Rostock
Rostock, Germany
petra.groeber@uni-rostock.de

Martina Weicht
University of Rostock, Germany
IT Science Center Ruegen, Germany
martina.weicht@uni-rostock.de

Abstract

Speech interfaces are an alternative way of interacting with a computer. Not only blind and visually impaired people could profit from this form of interaction, but anyone. In order to find out whether people actually want to talk to their computers and — most interestingly — how they want to talk to them, we conducted a Wizard of Oz experiment. We assumed that beginners are task-oriented and would use natural language while advanced users would interact in a menu-oriented and command-based fashion. We expected both groups to change their strategy to a more efficient one.
In this paper we present the setup, the conduction and the results of this user study describing how users dealt with *Talking to Computers*.

1 Introduction

Speech is a natural form of communication that is used in our everyday life. Since we learn to speak during the first years of our life, it seems obvious to use speech for the interaction with a computer. Moreover it enables us to interact with machines without the use of additional tools. We do, however, use keyboard, mouse and (touch) screen. So the question that arises before any speech-related research should start is:

How do people want to talk to their computers — and do they want to talk to them at all?

The question posed above originates from different areas of research: dialogue systems, alternative user interfaces and user/learner models. While the first two are directly related to the study on talking with computers, we feel that user/learner models are somewhat connected to it as well and need to be considered in future systems.

Interacting with a machine as if it was a human being is a long-cherished research objective. In the last years, especially in the field of telecommunication and call centres, many speech-enabled systems have come onto the market. Starting with simple command-based systems, the development of speech interfaces has been steadily increasing by means of naturalness during the last years. Contrarily, blind people are used to listen to their computers, but they do not talk to them (yet). One reason is that talking can be disturbing in situations where other people in the room are forced to listen like in open-plan offices, lectures or public libraries. From a technical point of view, language (and therefore speech interaction) is dynamic, fuzzy and interpretable. This leads to misunderstandings and unsatisfied users, who claim that the traditional way of interacting with a computer is more reliable.

To avoid these experiences, the design of a functional and effective speech processing system requires a careful analysis of the user's behaviour. The design of a functional and effective speech processing system requires a careful analysis of the user's behaviour. Therefore, we describe the setup and the results of a qualitative user study, which was realised as a Wizard of Oz experiment (WoZ) with both

sighted and visually impaired people, hoping to find clues for communication strategies for alternative user interfaces, especially for audio interfaces. In a WoZ experiment a human operator (wizard) is hidden behind a computer interface in order to simulate conversation with the user, who believes to be interacting with a fully automated prototype [4]. This reduces the development effort and allows the simulation of systems which are not yet implemented. The aim of our WoZ study is to determine the user's way of interaction with a machine, and to find out what it depends upon. This is the base for the identification of requirements which have to be fulfilled by a speech-based interface.

Section 2 will briefly introduce related work and background for the study. We define hypotheses to be evaluated during the study and describe the setup and the results of the study itself in section 3. Finally we will draw conclusions and provide an outlook on future work in section 4.

2 Related Work and Background

As research on dialogue systems and alternative user interfaces is not new, others have looked at these and related topics. This section describes *Wizard of Oz* tests in general and in the context of related work. Subsequently short introductions to our main research areas are given, which motivate the reasons for this study.

2.1 Wizard of Oz Experiments

In our study we do not want to test the usability of a system, but we want to find out how users interact with a solely voice-operated system. This means that interacting with the system should always deliver a correct problem solution in order not to influence the user's behaviour by errors which may result in unease. Therefore, the best method for our study is a Wizard of Oz experiment [12].

This type of experiment is a common method in the field of human-computer interaction. Subjects who interact with a system believe to interact with a fully functional application which actually is (partially) operated by a human being. Wizard of Oz experiments (WoZ) provide various information about the nature of the interaction and problems people will have with a system before it is actually implemented. Implemented systems could result in errors when the user acts correctly but unexpectedly. The "wizard", in contrast, is able to make sure, that the task the user wants to fulfil, will be solved successfully. This is the base for significant results.

Wizard of Oz experiments with similar goals have been conducted before. The study presented by Bennett and Edwards in [1] aimed at finding clues on how to present diagrams non-visually. Participants were confronted with circuit diagrams they were not allowed to see, but in order to explore them, they were allowed to ask questions about the diagrams. Bennett and Edwards found that, after a short orientation phase, participants explored the diagrams by moving from one element to another rather than exploring larger structures.

The aim of the work of Murao et al. in [16] is to generate dialogue examples which can be mapped to queries and search results in order to understand natural language requests and output natural language answers. Moreover it describes the realization of a framework for a WoZ which facilitates the performation of the wizard's tasks like interpreting user speech, executing searches and generating replies. Janarthanam and Lemon [10] conducted a WoZ that collects data about the use of *referring expressions* from users having different levels of expertise in a real situated task domain, i.e. setting up an Internet connection with a spoken dialogue system. The work of Gong and Lai in [14] describes a test on the different effects between natural and synthetic voices in voice user interfaces. The results show that purely synthetic text–to–speech systems perform better than mixed ones. The paper by Edlund et al. [6] focusses on the human-likeness of the user's responses as a measurement for the naturality of the overall system.

Carlson et al. [2] describe the differences between *interface metaphor* and *human metaphor* and state that human-like systems are not necessarily more efficient, but can profit from certain characteristics.

The WoZ study of Hanson et al. presented in [9] examines how elderly users navigate the Web using only speech. After a short introduction to the features of Internet Explorer and Firefox, the users were asked to navigate using only speech, whereas the experimenter controls mouse and keyboard. The authors learned that experienced users mastered the task, but less experienced users did not as they lack a deeper understanding of the functionality. They also found that people with less experience tend to use longer sentences to explain what they want to do rather than using short commands.

During Parente and Bishop's experiment in [18] the *wizard* is not hidden from the user but presented as an intelligent auditory display the user would directly interact with (along certain rules, of course). The research questions behind this experiment were slightly different than ours. The authors aimed at identifying problems with typical office tasks when using an auditory display, methods applied by the users to complete these tasks, and an operating mode that would prevent potential problems while best supporting the users' methods. The results of their study include problems with remembering information (for both longer and very short time periods), with a lack of visual clues (as a potential reason for not remembering information), and with identifying the next step in order to complete a task (combined with a liking of display initiated prompts). These findings were applied to *Clique*, the authors' prototype of an auditory display.

While many WoZ tests refer to the analysis of multimodality (including speech) in ubiquitous environments, we are interested in *how* users speak to their computers. Therefore, we conducted our own WoZ experiment.

2.2 Alternative User Interfaces

Alternative user interfaces aim at compensating deficiencies a user might have when using a computer. Hadley [7] includes both hardware and software in his discussion on alternative user interfaces. He names trackballs to be used instead of a computer mouse, large-print keyboards, eyetracking systems, screen readers, speech and gesture recognition, and the use of brain signals. Google advertises an alternative Google Reader user interface, providing a different graphical representation for its feed reader application[1].

While hardware solutions provide an alternative to other (potentially unusable) hardware and Google replaces a graphical interface by another graphical interface, screen readers and similar products "translate" graphical representations of screen contents into e.g. audio ones. Parente, however, says that "the goal of audio adaptation is not to mimic visual interfaces (...), an audio interface must do more than provide a superficial layer between user and screen." [17]. If, for example, the user writes an e-mail, he/she does not necessarily need a visual interface, Parente says, but a means to enter recipient, subject, and the message itself. These are subtasks to be performed in order to complete the e-mail writing task and as long as they are appropriately supported, the user will be able to perform the task.

So the ultimate goal of audio interfaces as one type of alternative user interfaces would not be to read the screen content, but to support the tasks a user wants to complete in audio – independently from any visual representation.

[1] Google launches an Alternative Google Reader User Interface: `http://www.socialtimes.com/2010/03/google-launches-an-alternative-google-reader-user-interface/?red=rb`, last checked April 27th, 2010.

2.3 User Models, Learner Models and Adaptation

Adaptation in application programmes distinguishes between *adaptability*, where the user adjusts parts of the user interface (explicitly), and *adaptivity*, an automatic (implicit) customization of the user interface performed by the system [5]. Computer systems, including eLearning applications, that are able to adapt themselves to their users or learners generally do so based on user models or learner models respectively. User Modeling collects information about the user of an application in some type of user profile or preferences. What type of information is collected depends on the application and task domain. Learner Modeling is mostly connected to Intelligent Tutoring Systems (ITS), which adapt the learning content to each individual learner. They usually (should) contain the learner's knowledge (e.g. as a subset of the expert's knowledge) and his/her skills [15].

Thinking about voice-operated systems, we assume that speech interaction with computers does not only need an adaption of the presented content, but also to the communication style of the user. In a sense voice-operated systems and maybe even application programmes in general should turn into (small-scale) ITS in order to provide more help and support to their users. An interesting question in this context is: Does the adaptation of the communication style rely on a user model or should it simply imitate the user's style?

2.4 Dialogue Systems

Carstensen [3] describes a dialogue as a communication with at least two persons, who perform and perceive utterances. Hamerich [8] adds a purpose and defines a dialogue as a goal-oriented speech-based interaction. When it is possible to reach this goal by talking to a speech interface, we speak of a dialogue system. In comparison to command-based systems, a dialogue system is able to process the user's requests in several correlating steps. As a dialogue consists of sequences of utterances with changing initiatives, a dialogue system must be capable of understanding the user's intention, accessing the backend for retrieving the required information and providing an appropriate reaction. The key component of a dialogue system is the dialogue manager. This module is responsible for the behaviour of the system. In this connection it is important that the user never feels restricted in his possibilities [13]. The required complexity is dependent on how users interact with a system, i.e. how they behave and what they expect. To model such a system according to the user's needs and wishes, it is necessary to investigate people's behaviour when interacting with a computer. Another interesting question which arises in this connection is whether users react in a specific way because they think this is the best way, or because they suppose their way to be the only possible.

The benefit of dialogue systems over conventional speech interfaces is that problems can be solved in cooperation with the user over several steps. Adaptive systems help to individually react according to the user's needs. Thus a dialogue system – as one form of alternative user interfaces – *can* be more efficient, easier to use and more convenient, depending on the task to solve. Nevertheless it should be mentioned that using speech also brings along difficulties. As long as systems are unable to fully interpret natural language, their usage may yield to disappointment when they do not react as expected. Since speech is unsharp and ambiguous, there are tasks where a distinct behaviour or formal languages perform better. However, our aim is to improve existing systems, as they can help disabled people and approach the goal of interacting with a machine without requiring additional tools.

Table 1: Two-dimensional matrix

		Mindset	
		menu-oriented	task-oriented
Phrasing	commands		
	natural language (full sentences)		

3 User Study

We conducted a Wizard of Oz experiment to find out how users deal with a voice-operated system and which strategies they use to handle it. We wanted to investigate how users with different background knowledge approach the task of solely using a speech interface to interact with a computer. The subjects neither had a screen, a keyboard nor a mouse. They were only provided with a microphone to talk to the computer and speakers to hear its responses. The following sections present our hypotheses, the setup of the study, its conduction and finally the results.

3.1 Hypotheses

In our study we examined whether there is a preferred or intuitively used type of interaction for a solely speech-based interface. Furthermore we scrutinized whether there are differences between experienced and inexperienced users, between blind or visually impaired and sighted users.

The base for our considerations is the two-dimensional matrix shown in table 1. The two dimensions are *mindset* and *phrasing* since we opine those to be the most important categories of perceivable behaviour. The mindset describes the mental way of how people proceed by using only speech to operate a computer system. The other dimension, phrasing, concerns how people formulate their utterances to instruct the computer system.

From our point of view, the dimension *mindset* can be divided into two categories – *menu-oriented* and *task-oriented*. Someone who regularly uses a computer and the Internet will have a mental idea, i.e. a complex mental model [11], of their usage. Therefore, when using only speech to operate the system, he will proceed in a menu-oriented way just like when using keyboard, mouse and screen. We assume users with broad experience in using graphical user interfaces and computer operations to be moving along menu bars to fulfill the given task. They will even use the exact words of the menu labels. Such users act in a *menu-oriented* fashion. Users with less experience, however, do not have such a clear idea of applications, so we assume they rather will proceed in a *task-oriented* way. Due to less experience with graphical user interfaces they do not have such detailed knowledge of their structures. We suppose they will focus on the task and use more complex instructions rather than formulating every single step like "translating" the usual way of using the computer.

The second dimension, *phrasing*, describes how people formulate their utterances to instruct the computer. It can also be divided into two categories – *commands* and *natural language (full sentences)*. Commands are instructions formulated as short utterances which are not sentences, i.e. imperatives or nouns. Natural language instead is characterised by full sentences and reminds of a communication between humans – complex instructions formulated in full sentences including phrases of civility and filler words.

As shown in the matrix in table 1 these respective categories form four interaction types:

- menu-oriented / commands
- menu-oriented / natural language
- task-oriented / commands
- task-oriented / natural language

Due to our preliminary considerations the following hypotheses emerged:

1. Beginners interact in a task-oriented way and speak naturally with a solely voice-operated system.
2. Advanced users interact in a menu-oriented and command-based fashion with a solely voice-operated system.
3. Both user groups will change their strategy of communication into a task-oriented and command-based interaction in order to achieve more efficiency.

We assume a strong correlation between menu-oriented mindset and command-based phrasing. This assumption is based on our consideration that people with a broad experience will be moving along the menu bar and will even use the exact menu labels to interact with the speech interface. Because of their experience and the common desire to fulfil a computer task efficiently, they will use simple commands. Besides, experienced people are aware of the capabilities of a computer and the fact that a computer is just a machine, so they do not believe a speech interface could work in any other way except for the one they know.

We also suppose a correlation between a task-oriented mindset and the use of natural language. As people with less experience will – due to their inexperience – not remember the exact paths and the exact menu labels to fulfill a given task, they will supposedly use more complex instructions. Their ingenuousness in technology will lead them to use a more ”human” communication style – full sentences with phrases of civility and filler words – instead of plain commands.

Furthermore we suppose a transition from both menu-oriented/commands and task-oriented/natural language to task-oriented/commands which we assume to be the most effective way of communication. Hence a user will change his/her way of communication to this type in order to fulfill tasks as fast as possible. Task-oriented instructions are more effective than menu-oriented ones because not every single step needs to be mentioned and the communication partner derives intermediate steps. Since commands usually are shorter than natural language, they tend to be more effective. They are, however, not as expressive as full sentences. A user who interacts in natural language first will probably omit phrases of civility and filler words during repeated usage, as this reduces the time needed while leading to the same results.

3.2 Setup

The Wizard of Oz experiment sketched in figure 1 took place in two different rooms. The test candidate, sitting in one room, was equipped with a microphone to talk to the computer and with speakers to hear the system’s responses. He had no screen and thus no graphical user interface. Furthermore, he had no input devices to point at or type in anything. The audio devices were connected to a computer running a speech synthesis programme. It was connected to the wizard’s computer via IP telephony (Skype). This allowed the latter to hear the test candidate’s spoken instructions. The system’s responses were

sent back to the candidate as synthesised speech[2] issued by the wizard. In order to create and maintain the impression of a computer system talking to the test candidate, we decided not to let the wizard talk himself. Instead we developed an application which allows the management and synthesis of speech utterances based on the *Microsoft SAPI*. This application was controlled by the wizard and enabled him to send predefined prompts, as well as to create prompts on the fly by typing his replies on the spot. The audio data (the user's instructions and system responses) was recorded using a regular voice recorder (and the Skype plugin Pamela[3] as a backup) for later transcription and analysis.

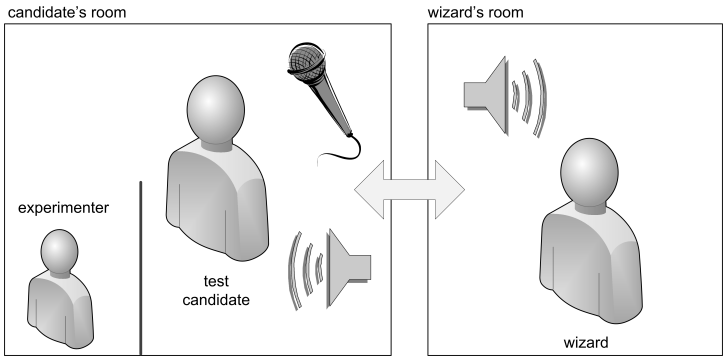

Figure 1: Wizard of Oz test

Before inviting test candidates to participate in our study, we conducted a pre-test with three colleagues which provided valuable feedback for the "real" test.

In order to collect both demographic information about our participants and their impression on the scenario, we prepared two questionnaires. The first questionnaire contained demographic questions like age and gender of the participants as well as questions about their vision, and their computer and Internet experience. This questionnaire was filled in before the actual test during a conversation between one of the experimenters and the participant so as to get to know each participant and to let him/her adapt to the situation.

The second questionnaire, too, had the character of an interview. We asked the participants how they felt using the speech interface without having any visual feedback, whether they would use such a system in everyday life and in which situations they could imagine to do so. Furthermore we asked whether the system was reacting in the expected manner and how they would wish the system to react – short answers or more like a "human" conversation.

3.3 Conduction

For this qualitative study we acquired a total of 18 participants. Three of them were recommended by the association for blind and visually impaired people in Mecklenburg-Western Pomerania. Three other subjects attend the *Seniorenakademie*, a lecture programme for senior citizens at the University of Rostock. We also contacted Rostock's adult education centre in order to include computer class beginners. Unfortunately, classes had not started again after winter break, so only one person, the director of the centre, participated in our study. The remaining subjects were colleagues, researchers and lecturers at the University of Rostock, or friends of the authors.

The majority of tests took place at the University of Rostock on two days in late March 2010. Three test sessions with car drivers — "situation–blind peple" — were scheduled for car trips between the

[2] ScanSoft Steffi

[3] http://www.pamela.biz/en/

university and the IT Science Center Ruegen. The tests themselves took between 15 and 30 minutes per participant.

Altogether eleven male and seven female participants participated in our study. We classified three age groups: seven participants were between 20 and 40 years of age, another seven participants ranged between 41 and 60 years and four participants were older than 60 (three of them even older than 70) years.

Most participants (14) had "regular" vision, some of them wore glasses but none had a visual impairment. Out of the four participants with low vision, one used a screen magnifier as an assistive device. Another participant used the computer through her husband, who would read e.g. e-mail to her and type in what she wished to reply. The rest did not use any assistive technology but two out of all participants (one with vision impairment and one elderly user) rely on low screen resolution, making fonts and icons on the screen appear as large as possible.

One question of the questionnaire asked participants to rate their own computer experience. Two participants considered themselves to be beginners, ten to be advanced users and six rated themselves as experts. Most of our subjects have been using computers for more than ten years and the Internet for several years as well.

After the first questionnaire, participants were introduced to the situation of sitting in front of a computer with an Internet connection provided and all common applications installed. The only difference to a common workplace was that there was neither a screen, a keyboard nor a mouse but a microphone and speakers. Then the subject was asked to solve two given tasks. We performed two tasks in order to let the subject get used to the situation. Besides we assumed to observe a change in system handling due to the user's adaptation towards it over time. In both tasks the subjects were asked to check their e-mail. They received exactly one mail which again contained a task on performing an online search. The search result should be sent back to the e-mail sender. The first task was to investigate the price of a book whereas the second task was to look up the weather forecast. The subject was free in how to solve the tasks while the wizard tried to simulate the system the user expects to have. Both menu-oriented and task-oriented approaches were possible.

3.4 Results

For a better interpretation we transcribed the recorded audio files. Since an interpretation of written words is very subjective, we used the following method. First each of the authors interpreted and classified the transcribed material individually. Afterwards all three of us met to discuss the results. In most cases we agreed, in some cases the classification was accompanied by a discussion. In this section we present the results of our study, first the classification of the subjects into interaction types, then other interesting results the study yields.

3.4.1 Interaction types

While evaluating the data and trying to classify the users with the help of the matrix described in section 3.1, we realised the need to extend it by one category for each of the two dimensions. The two categories for *mindset* – menu-orientation and task-orientation – were not sufficient. Since one subject (TP11) clearly interacted in a task-oriented way, it was somewhat ambiguous to classify the others. Some subjects were not clearly menu-oriented but were obviously familiar with the used applications, that means they had a clear idea of the graphical interface but did not move along the exact menu structure – we called this *application-oriented*.

The two categories of the dimension *phrasing* were not sufficient either. Some subjects clearly used commands – imperatives as one-word instructions. Others used natural language, i.e. they uttered full

sentences including phrases of civility and filler words. Yet others formulated their instructions somewhat inbetween these two classes. They primarily used groups of words as if they had in mind what would happen on-screen and just formulate every mouse click they would do – we call this *phrased commands*. Both the new categories and the classification of the subjects are shown in figure 2.

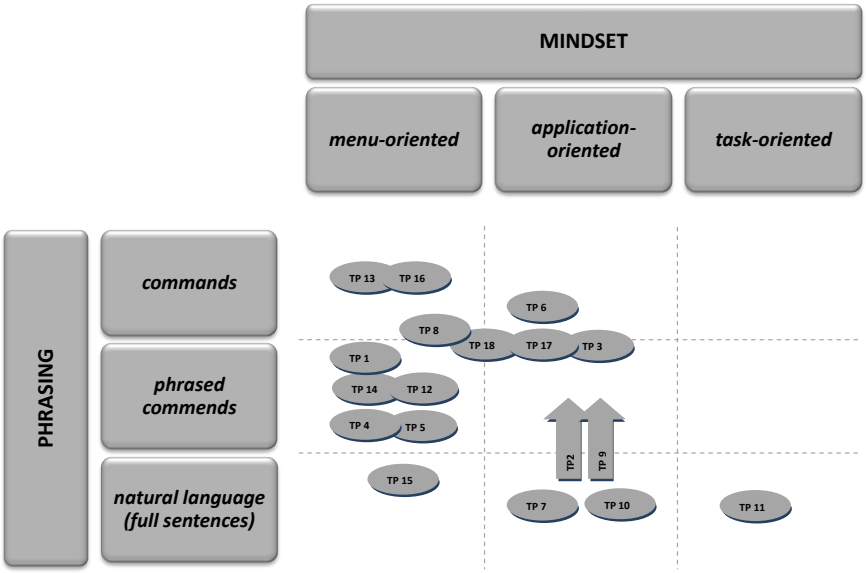

Figure 2: Results

As shown in our new matrix, with the new categories for each dimension, we now have nine interaction types:

1. menu-oriented / commands
2. menu-oriented / phrased commands
3. menu-oriented / natural language
4. application-oriented / commands
5. application-oriented / phrased commands
6. application-oriented / natural language
7. task-oriented / commands
8. task-oriented / phrased commands
9. task-oriented / natural language

People who are using the exact words from menu labels and who are acting in the corresponding steps, like opening the e-mail programme by selecting *"new mail"* from the menu bar and focussing the subject text field just as they would do by using the mouse, interact in a *menu-oriented* fashion. They have profound knowledge of graphical user interfaces and computer operations. People who care about starting and quitting programmes without using the menu are called *application-oriented* users. People who are detached from the graphical interface and who are just focused on the task to fulfil, interact in a *task-oriented* way. The reference to graphical user interfaces (GUI) is strongest in the case of menu-oriented interaction and smallest in the case of task-oriented interaction. Along with the decreasing dependence from the GUI, the complexity by means of task-aggregation and thus simplification of the user's workload increases. We consider a task-oriented mindset to be the most complex, most natural and most effective form of interacting with the computer by voice.

Apart from the mindset, also the *phrasing* is an important description of the user's behaviour. The simplest form are short *commands*, i.e. imperatives or nouns like *"close"*, *"send and receive"* or *"Google"*. The second form of expression is *phrased commands*. These are also imperatives but expressed as part sentences like *"open browser"*, *"check mails"* or *"write new mail"*. The most natural form are full sentences or colloquial expressions like *"Please write a new mail to Gabi"* or *"Leave the mail"*. This phrasing-class is called *natural language*.

3.4.2 Observations and Examples

In our study the interaction types *task-oriented/commands* and *task-oriented/phrased commands* did not occur at all. The *task-oriented/natural language* type occurred once (TP11). This subject was totally detached from the graphical interface and basically fulfilled the tasks in just one sentence each: *"Computer, Internet, check weather in Munich for the weekend and mail the forecast to Gabi"*.

Another interesting observation is the behaviour of TP2 and TP9. Both were first classified as *application-oriented/natural language* but changed to *application oriented/phrased commands* for the second task. It seemed both needed time to become familiar with the system and then used a more efficient way to interact. TP2, for example, first used full sentences like *"I want to [...] look up the price of the book"* or *"Thanks, then leave again"* but in the second task switched over to a list of commands like *"Put in password, check e-mail, new e-mail"*. Similar observations were seen with TP9. In the first run, TP9 used phrases of civility such as *"please"* and *"thanks"* several times, e.g. *"Please read first e-mail"*, *"Ok, thank you"* and *"Thanks. Well. Open Browser."*. He somehow talked to the computer as if it was a human being: *"Do you have results?"*. In the second run TP9 left out phrases of civility and used short commands like *"Open e-mail programme"*, *"Do I have new e-mails"* or *"Open browser"*.

Most subjects can clearly be classified into one of the fields of our matrix. As supposed, most users interacted in a menu-oriented way (or at least in an application-oriented way if they are not too attached to the menu structure) but have a clear idea of the on-screen content.

Sometimes a precise distinction between the categories *commands* and *phrased commands* was not easy. Subjects TP13 and TP16 clearly interacted in a command-based fashion as they used one-word instructions like *"Firefox"*, *"E-Mail"*, *"Send"*, *"Books"* or *"Vendor"* whereas TP1, TP4, TP5, TP12 and TP14 used phrased commands like *"Send e-mail"*, *"open new e-mail"*, *"please send"*, *"please show price"*, *"open first entry"* or *"open a new tab"*. TP8 was placed inbetween these two categories as this subject changed categories during performation of each task. TP8 mostly used phrased commands, but as soon as he got to the point to fill in forms like an e-mail form he used one-word instructions like *"Address"*, *"Subject"* and *"Text"*. In these situations the subject expects a response of the system to be ready, just like the cursor would blink for a visual feedback to be ready to type.

Also TP15 plays a special role since this subject was actually telling what she would be doing: *"I want to login [...]"*, *"Then I read [...]"*, *"Then I put in [...]"* or *"Then I close [...]"*. She had more difficulties than other participants in instructing the computer and rather recited her own (graphical) interation with computers.

3.4.3 Tendencies and Generalisation

The second questionnaire showed that most of the subjects could indeed imagine to use a speech-operated system in their everyday life. Asking for situations in which they could imagine to do so, they said: checking e-mails while driving a car, looking for a recipe while cooking or in general to search the Internet while their hands are not free. Surprisingly most subjects thought the system used for the study was a real, existing system and were astounded how well it worked. They were satisfied with the reactions of the system and stated that it behaved just like they had expected. Asking for how they wished the

system to react, the answers vary. Some subjects, mostly the elderly, would like to experience a "human like" reaction — they even would like to be asked "How are you?" — whereas other users, mostly experienced ones who use a computer very often (and professionally), would prefer short answers – only containing the desired information.

We also analysed the data by means of patterns. First of all, we analysed the correlation between the phrasing and mindset categories with an *association rule learner* algorithm, which is implemented in the data mining tool *KNIME*[4]. We observed the following trends:

- menu-oriented users tend to use phrased or single commands and vice versa
- all people who interact in a task-oriented way use natural language
- all people using phrased or single commands interact in a menu- or application-oriented fashion
- people using single commands or phrased commands never interact in a task-oriented way
- task-oriented people never use commands

These results prove that menu-orientation is strongly connected with command-based formulation. The formulation of more complex structures comes along with task-orientation. This is fairly obvious as formulating complex tasks can better be realised using full sentences instead of short commands. This is why single commands and task-orientation exclude each other. Moreover the formulation of full sentences in connection with a menu-oriented mindset is rarely used as it is rather ineffective.

In a second step we tried to find relations between the target attributes and the answers of the first survey (e.g. computer expertise, visual faculty). One of the hypotheses was that people who do not have a visual impairment and who have been using computers for several years, are more likely to have a menu-oriented mindset and to use command-based phrasing, than people with less computer experience. This hypothesis was not confirmed. It rather seemed that people have a particular idea of how a computer interface works in general, since the usage of a computer has become an integral part of our society. People have a well-practised method in mind, using a computer with mouse, keyboard and screen, which they transfer to the speech-operated system. When they were told, they could have used a different type of interaction – for example a task-oriented approach – most subjects said they did not think of such an idea at all. This indicates that people are rather set in their way to deal with a computer system. On the other hand, experienced subjects did not trust in a speech interface working with natural language because they already gained negative experience in using those systems.

Another interesting aspect was that some subjects replied to their mail before executing the task. Afterwards they opened the mail again to answer it by communicating the search result. This behaviour is different from the GUI-based interaction and might be explained with the untrained acoustical multitasking skill. People are used to GUIs and are able to manage several open windows at a time. The acoustic memory seems to be smaller or shorter, i.e. acoustic information is forgotten faster than visual information. With a screen, a user has the whole information at a glance, whereas the speech interface requires sequential processing. Another reason might be that the user is in an unfamiliar situation where he is not able to foresee the next step as he is not used to the system.

We also observed that the more uneasy the person feels, the more natural the utterances are going to be, i.e. natural language increases with unease. But also complex tasks can lead to a change in interaction types. We noticed that when the subjects enquired about search results, it was such a cognitive effort that they did not find a menu-structured way to solve the problem.

[4] http://www.knime.org/

They switched over to a task-oriented way, like TP 3 did in the following extract of his utterances:

***User:* to inbox**
Wizard: programme opened
...
***U:* read mail**
W: Sender: Gabi, Subject: Gift for Ben. Hello Katie[5], do you already have a gift for Ben? [omitted]
***U:* answer**
W: ok
***U:* I'm going to look for the price**
W: text recorded
***U:* logout**
W: confirmed
***U:* Internet**
W: the browser is now open
***U:* weltbild.de**
W: address loaded
***U:* I am looking for the price of the book "The Wizard of Oz"**
W: The book "The Wizard of Oz" costs 13 Euros
***U:* ok, thanks**
W: you're welcome
***U:* back to the inbox**
...

3.5 Critical Discussion of the Study Setup

The objective of our study was to find out whether there is any preferred or intuitively used type of interaction for a solely speech-based interface. A WoZ appears to be the most promising method for that. Since we are interested in the dialogues themselves and in verbal communication in general, we employed this qualitative method. It does, however, imply a complex evaluation since the recordings need to be transcribed manually and every single phrase and word uttered by the subjects needs to be analysed. This is why we limited our study to 18 subjects.

We assumed to find differences in speech-based computer interaction between sighted and visually impaired people as well as between beginners and advanced users. Therefore we decided to cover a wide range of participants – beginners, advanced users, experts, sighted and visually impaired as well as blind subjects. Unfortunately we only found four visually impaired subjects. It also appears to be difficult to find absolute computer beginners as today most people have already been in contact with computers in some way.

One major source of errors was the wizard. We tried to bring in consistency by always using the same person as wizard. Still it was hard for the wizard to always react in the right manner. His strategy was to try to imitate the subject's interaction strategy, i.e. if the subject used commands, the wizard would try to provide short answers as well. This turned out to be quite difficult with regard to reaction time, since in many cases the wizard still needed to type in his answers. If a (phrased) answer took too long, the subject would be impatient and the system would make an unfinished impression. Also we assumed less advanced and in particular elderly subjects to talk to the system as they would talk to a real

[5]Name changed by the authors.

person. Therefore we refrained from letting the wizard react computerlike – using commands and short sentences –, in order not to influence subjects by reacting in a typical machine-like manner.

4 Conclusion and Future Work

In this paper we have examined how users interact with computers using speech only. The aim of this qualitative study was to evaluate how dialogue systems should be designed to provide a high usability. This decreases the effort of development and allows an optimized understanding of natural language. The results of this study illustrate that there are strong relations between phrasing and mindset. It has been shown that most people – independent from age, gender, experience or profession – stick to the GUI. This can be explained by a kind of mental conditioning: people are just used to this form of interaction. Surprisingly even visually impaired people have shown this behaviour. Moreover we have learned, that most of the participants categorise natural language as helpful. It was interesting to see, that in case of unease or in the need to solve a complex task, users switched to natural language. Furthermore, when explicitly offering the possibility to naturally interact with the computer, they were surprised and stated that this would be much easier than application- and command-oriented control. We suppose that many people are virtually "blind" towards this form of interaction because they do not believe it to work. Besides they are simply familiar with GUI interaction.

As the authors of this paper, we have approached the idea of talking to a computer with different motivations:

- Dialogue systems have to be modelled according to the users' needs. This results in the question, how people today interact with speech interfaces and which problems occur in doing so. Another interesting point is, whether users think they are forced to act in the specific way they used in order to make themselves understood by the computer.
- What alternative strategies can be provided e.g. to blind and visually impaired people to confront them with highly vision-oriented media and contents like websites? Websites in particular are optimized for mouse navigation and thus primarily designed for sighted people. Since the Internet becomes more and more important in our society, blind and visually impaired users should be able to have access to online information.
- (How) Can speech help in mobile learning settings where the learner is "temporarily blind"? In situations when he focusses his attention on other objects than a computer screen, speech input and output seems to be the obvious choice – one that does not only concern eLearning systems but any other application programme.
- How do computer systems that respond to fuzzy and interpretable speech input, instead of distinct mouse clicks, need to adapt to their users?
- What type of help and support is needed in voice-operated systems?

These questions could not fully be answered. An interesting outcome was the fact that users are strictly GUI-oriented and rarely think of acting naturally. This is why we plan a second user study with only menu- or application-oriented test candidates. They will be seperated in two groups, where the first group is told about the possibility to interact with natural language. Afterwards the learned interaction types will be compared and related to ease of use, wellfare and effectiveness. Furthermore we plan to design a system which adapts to the user, depending on the form of phrasing he decides to use. Another open question is whether an adaptation is only required between different users or also over time. These topics will be addressed in our future work.

Acknowledgements

We would like to thank everyone who volunteered in talking with computers. Thank you for letting us spy on how you did it, for your thoughts on the topic and for your advice.

References

[1] D. J. Bennett and A. D. N. Edwards. Exploration of non-seen diagrams. In *Proceedings of the 5th International Conference on Auditory Display (ICAD)*, 1998.

[2] Rolf Carlson, Jens Edlund, David House, Mattias Heldner, Anna Hjalmarsson, and Gabriel Skantze. Towards human-like behaviour in spoken dialog systems.

[3] Kai-Uwe Carstensen, Christian Ebert, Susanne Jekat, Cornelia Ebert, Hagen Langer, and Ralf Klabunde. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Springer, 2009.

[4] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of oz studies: why and how. In *IUI '93: Proceedings of the 1st international conference on Intelligent user interfaces*, pages 193–200, New York, NY, USA, 1993. ACM.

[5] Alan Dix, Janet E. Finlay, Gregory D. Abowd, and Russell Beale. *Human-Computer Interaction*. Prentice Hall; 3rd edition, 2003.

[6] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Commun.*, 50(8-9):630–645, 2008.

[7] Daniel Andrew Hadley. Alternative user interfaces, 1999.

[8] Stefan Hamerich. *Sprachbedienung im Automobil*. Springer, 1st Edition, 2009.

[9] Vicki L. Hanson, John T. Richards, and Chin Chin Lee. Web access for older adults: voice browsing? In *UAHCI'07: Proceedings of the 4th international conference on Universal access in human computer interaction*, pages 904–913, Berlin, Heidelberg, 2007. Springer-Verlag.

[10] Srinivasan Janarthanam and Oliver Lemon. A wizard-of-oz environment to study referring expression generation in a situated spoken dialogue task. In *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation*, pages 94–97, 2009.

[11] P. N. Johnson-Laird, Vittorio Girotto, and Paolo Legrenzi. Mental models: a gentle guide for outsiders, 1998.

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

[13] Uwe Koloska and Matthias Pohl. Modellbasierte generierung von sprachdialogen für eingebettete systeme. In *20. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, September 2009.

[14] Jennifer Lai Li Gong. Shall we mix synthetic speech and human speech? impact on users performance, perception, and attitude, 2001.

[15] Alke Martens. *Ein Tutoring Prozess Modell für fallbasierte Intelligente Tutoringsysteme. DISKI 281*. Akademische Verlagsgesellschaft mbH AKA, infix, Berlin, 2004.

[16] Hiroya Murao, Nobuo Kawaguchi, Shigeki Matsubara, Yukiko Yamaguchi, and Yasuyoshi Inagaki. Example-based spoken dialogue system using woz system log. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, July 2003.

[17] Peter Parente. Clique: a conversant, task-based audio display for gui applications. *SIGACCESS Access. Comput.*, (84):34–37, 2006.

[18] Peter Parente and Gary Bishop. Out from behind the curtain: learning from a human auditory display. In Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson, and Saul Greenberg, editors, *CHI Extended Abstracts*, pages 2575–2584. ACM, 2009.