User expectations and selected notions in the context of the spoken dialogue system INSPIRE

Rosa Pegam[1], Jan Krebber[2]

[1]Fachbereich Kommunikationswissenschaft, Universität Duisburg-Essen, Germany
Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
rosa.pegam@gmx.de
jan.krebber@rub.de

Abstract

This paper investigates the interaction with the INSPIRE spoken dialogue system as perceived by users from a communication-theoretic perspective. We found a practical interest in research into user expectations in the context of interaction with spoken dialogue systems. When designers of these systems create the various dialogue modules they usually rely on designs of previous interactive speech systems. This is not always the most recommendable way to deal with the matter because in many cases the earlier reference systems (e.g. automatic telephonic train information services) were designed for purposes different from the current ones (e.g. voice controlled smart-home applications). This is certain to have a negative effect on the eventual performance quality of the newer system because the dialogue structures were not developed exclusively for the system in question. This paper investigates expectations that users build up while interacting with the INSPIRE system.

The objective of this paper is to reveal the most striking kinds of expectations and those most relevant as to the dialogue flow. Furthermore, a categorisation of system limitations regarding dialogue shall be established. We assume that when user expectations are known to the system developer, this knowledge can be applied during the system implementation and will lead to a reduction of ill-structured dialogues (i.e. incomprehensible structures).

The interdisciplinary approach taken in this paper curls up challenges as regards thematic commitment: the main focus of this paper is on the actual interaction events between man and machine. This paper is strongly empirically oriented. Although literature treating the subject of user expectations regarding dialogue systems has been published, the vast majority of these texts regard dialogue systems based on modalities other than voice-controlled user interfaces (e.g. graphical user interfaces). This circumstance makes a practical approach to the subject necessary, that is, a study of the subject in very close connection with the INSPIRE system tapping the wealth of experience the authors have collected while assisting with the development of the system prototype at the Institute of Communication Acoustics Ruhr-University of Bochum during 2003 and 2004.

1. Introduction

Spoken Dialogue Systems (SDS) are software applications with vocal user interfaces, that is, they allow for two-way spoken interaction between humans and machines. It is the aim of such systems to facilitate convenient interaction between human users and the various systems. Over the last few years spoken dialogue systems have been used within a wide range of domains and have now reached a relatively sophisticated stage. SDS are designed for various kinds of users: experts use SDS (e.g. in medical environments where doctors use dictation systems for fast recording of information), in technical environments (e.g. motor mechanics use dictation systems for fast and hands-free recording of technical defects while inspecting vehicles). Novices use SDS for instance for call routing or travel inquiry applications, cinema-related issues or for retrieving information about eating out. This is just a small selection of instances of the actual variety of application areas.

Currently telephonic information services constitute the application area in which spoken dialogue technology is used most frequently. These kind of dialogue systems also seemed to be the most familiar to the test subjects who participated in the INSPIRE experiments. Most systems, even if they have reached a rather advanced level of operability, are still highly limited in capacity, i.e. their dialogic structures are very rigid and simple. Apart from those listed above, there are also more complex applications such as dialogue systems as part of so-called smart-home environments where novice users can voice-control devices in the house.

SDS with intuitive interfaces that should make them easy to use are difficult to design. Especially when the systems are supposed to be designed in such a way that users should be able to operate them without previous knowledge to trigger the desired actions. A crucial subject is the question of input modality, meaning how (by what means) the system should be operated. In this context we distinguish uni-modal (by keyboard-input only or by voice only) and multi-modal (a combination of different input methods, e.g. vocal input associated with touch screens and the like) systems.

In the framework of the EC-funded IST project INSPIRE (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces), a spoken dialogue system has been set up to control domestic devices in a home environment. With INSPIRE, one may operate different home appliances (blinds, fan, lamps, TV, VCR, electronic programme guide, answering machine) via natural speech. INSPIRE uses a state of the art Automatic Speech Recogniser (ASR), speech understanding and dialogue management technologies, and spoken output via pre-recorded or synthesized speech. The set-up of the system is described in Section 2. Section 3 explains the dialogue model, section 4 shows the strategies in dialogue management within the INSPIRE system. Section 5 describes the system improvements of the spoken dialogue system, section 6 explains the experiments conducted within the INSPIRE project, section 7 shows the communicative situation related to INSPIRE with technical and linguistic limitations. The paper closes with a conclusion and references section.

2. The Inspire spoken dialogue system

2.1. Who is it for?

INSPIRE was designed with a view to suiting novice users, that is, people who do need neither any previous experience regarding the operation of INSPIRE, nor experience with spoken dialogue systems in general. The system supports the technically inclined by simplifying the control of electronic devices. As mentioned above, the fact that the system should be usable without any previous experience or user instruction creates some obstacles regarding the system design: the system must have the ability to understand the most frequently used terms as keywords and there should be system help messages to assist the users in awkward situations. It must be possible to trigger the built-in system help at every point in the dialogue such as TV-related appliances or the answering machine. Furthermore, it makes life easier by making it possible for its users to remain seated and is therefore intended to contribute to the creation of a relaxed atmosphere at home. It can moreover be used as a valuable technical aid for the physically impaired, i.e. elderly or handicapped people who have difficulties in moving.

2.2. What is it for?

The INSPIRE spoken dialogue system was designed as an experimental prototype. One of its main features is that it provides a unique interface for different domestic devices of different grades of complexity such as TV/VCR combination, answering machine, lights, blinds and fan. In addition, the system may be operated in different situations, e.g. at home or from a remote location. To support inexperienced users, the system is designed for *natural* spoken language interaction, including sophisticated spoken help prompts. The mixed initiative dialogue structure is the same for all devices. In this way, the dialogue system has always the same hear-and-feel, even when operating complex devices.

The dialogue structure is single-task oriented. This means, INSPIRE is able to operate only one device and to perform one action at a time. This approach was chosen to keep a clear dialogue structure especially for the complex devices.

Another important feature is the ability of the system to propose potential actions the system can perform, in case of confusion. If the dialogue management module receives contradictory input values which do not lead to an action the system can perform, the dialogue manager provides different correct solutions to choose from by relaxing one or two constraints.

2.3. System Design

The INSPIRE dialogue system is built in a modular way. It allows easy adaptation to new environments, e.g. to replace the ASR, to use different speech synthesizers, or different hardware controls over the devices. The major system modules are shown in Figure 1. They are all grouped around the core of the dialogue system, the dialogue manager.

The acoustic signal is picked up by the microphone array. It contains several microphones which deliver audio signals in parallel into the *beam forming module*. The beam forming module focuses on the active talker and removes the other sources which are not in the beam, producing just one audio signal stream. This stream is sent to the *noise reduction module* which separates the desired signal of the talker from the unwanted background noise which was left within the beam of the beam former.

After this pre-processing, the audio signal is split to the *speaker identification / speaker verification module (SI/SV)* and the *ASR*. The speaker identification is essential for separating commands originating from valid talkers from those who are not allowed to use the dialogue system inside the house. The speaker verification is necessary for remote access by telephone to allow access to the dialogue system only to permitted persons. The decision of the SI/SV module is given to the *dialogue manager* which decides how to continue in the dialogue.

As an outcome of a recognition test performed during the system set-up phase (Trutnev and Rajman, 2004) a commercial ASR was chosen for the INSPIRE project. The ASR allows some editing in the acoustic model but no training towards a specific talker. After the ASR has detected an utterance, it closes after a certain time of silence and produces a character string. The ASR is disabled while the dialogue system is processing or giving speech output; which means that barge-in is not possible.

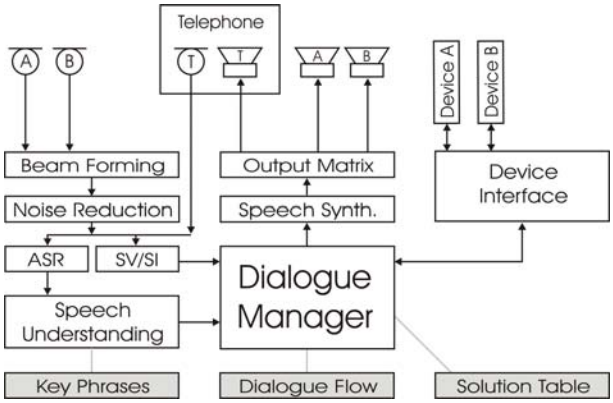


Figure 1: Block diagram of the INSPIRE spoken dialogue system.

The character string produced by the ASR is processed in the *speech understanding module.* Here, all the related utterances of the user are linked to canonical values which finally fill the open slots, i.e. attributes which enable the dialogue manger to come to a solution. The database for the speech understanding module is the *key-phrase* list. By making use of longer key-phrases instead of single keywords, the ambiguity can be reduced.

The *dialogue manager* interprets the incoming information of the SV/SI module and the speech understanding module. Beside of that, it needs information about the possible solutions which are stored in a solution table. Depending on the input information, the dialogue manager may select different dialogue strategies which are listed in a *dialogue flow* database. The *solution table* contains static information such as possible conditions of the devices, as well as dynamic information like the TV programme and the current condition of each device.

Once the next step in the dialogue is determined, spoken output can be generated. For example, the system asks the user for more information as long as it has not been able to generate a single solution. This speech output is generated by

a *speech synthesizer* and given to an *output matrix*. The *output matrix* routes the signal to the loudspeaker specified for this certain interaction.

When the dialogue manager comes to a valid solution, it operates the related device via the *device interface*. This interface is designed in a bi-directional way and allows a feedback from the *device*, e.g. in case there is a broken bulb in the lamp.

2.4. Wizard-of-Oz environment

In order to allow maximum performance of the ASR component, this module had to be replaced by a simulation which allows different speech recognition performances to be generated. This simulation was implemented in a Wizard-of-Oz (WoZ) paradigm.

A WoZ test allows one or more modules of a system to be assessed under controlled simulation conditions (Bernsen et. al., 1998). It is mainly used in case when respective modules do not exist or they do not perform in the desired way, e.g. during the system set-up phase. These modules are replaced by a wizard, which has to act like the replaced modules. This requires knowledge about the specifications and the behaviour of these modules.

In case of the INSPIRE system, the WoZ paradigm is used for system component assessment. A WoZ environment was build in a modular way to assess nearly all modules or databases individually. For the assessment of the minimum word error rate being allowed to gain an acceptable quality, only parts of the speech input modules were replaced, as shown in Figure 2.

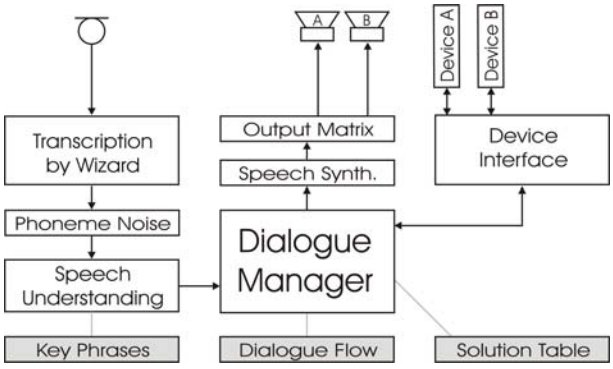


Figure 2: Set-up for Wizard-of-Oz tests.

The microphone array was replaced by a single, high quality microphone. This microphone allows full "acoustic observation" of the test room. The beam forming and noise reduction modules were not used as their performance influences the word accuracy, which should be controllable for a given purpose.

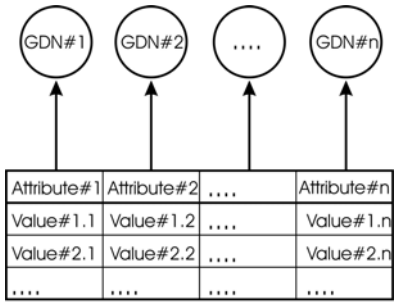
The ASR is replaced by a transcribing wizard. The wizard was capable of transcribing the utterances at least in the same periods of time that the ASR would have needed for processing the same utterance. The wizard transcribed literally everything the test participants were saying during the duration of the test. The wizard was controlled by a supervisor, which ensured transcription accuracy close to 100%.

The transcribed utterance is then transferred to a *phoneme noise* generator module (Trutnev et. al., 2004). Here, the incoming utterance is "noised" in a specified way, i.e. parts of the utterance are changed or left out on the basis of phoneme similarities. The relative probabilities of the substitution and deletion rates were estimated in a prior test with the commercial recogniser. The absolute probabilities were calculated from the parameters for the word accuracy, the word error rate and the substitution rate, defined for each test run. As the input word accuracy is nearly 100%, the output word accuracy is defined by setting the mentioned corresponding parameters.

To simplify the test set-up, the SV/SI module was left out, as during the tests only one test subject was placed in the test room, without additional noise sources.

3. Dialogue model in INSPIRE

The interaction with the system is task-oriented: system and user aim at completing a task. For the system each task is divided into a fixed number of sub-tasks. If the task is to record a movie, then the corresponding sub-tasks will be to define the time and day of the broadcast as well as the show-type (e.g. movie, documentary etc.). The tasks and sub-tasks which must be carried out are to be seen as a set of frames in which the fields are associated with attributes (e.g. "device" or "action") and values (e.g. "fan" or "switch on"). The dialogue model is a set of interconnected generic dialogue nodes (GDNs) which are processed by the dialogue manager. The dialogue manager follows the branching logic until an utterance can be produced. The GDNs determine the system's focus of attention, i.e. the expectations regarding user input.



Attribute	Device	Action
Value	fan	switch_on

Figure 3: GDNs as associated with the attribute-value pairs, according to Rajman (2004) and attribute-value matrix sample.

As shown in Figure 3, each GDN is associated with an attribute value pair. In order to fill in the obligatory fields, i.e. to determine a value for each attribute, the system must interact with the user. There are static and dynamic GDNs. Static GDNs are connected with static fields, i.e. values that are relatively stable such as the names for particular devices (e.g. "fan", "table lamp" etc.). Dynamic GDNs are connected with dynamic fields, i.e. fields whose values change frequently such as a list of movies which the system can record. Names and show times of movies change on a daily basis and must therefore be updated constantly for correct

retrieval of information. In general the dialogue model consists of two parts:

1. The specification of the necessary GDNs and
2. Choosing a strategy for proper dialogue management

4. Strategies in dialogue management in INSPIRE

The dialogue management module manages the course of the dialogue between the user and the system. At this stage decisions about which strategy to employ at which point of the interaction are taken. The dialogue management module follows predefined logical branches thereby activating the single GDNs and the branching logical steps, and collects the amount of information which is necessary for the generation of a solution. This will subsequently lead to either the performance of an action (e.g. switch on the fan) or to the generation of spoken output. The necessary pieces of information are collected by asking the user questions until enough slots have been filled for the generation of a solution. The strategies known to the dialogue manager are listed below:

Assertion (OK) The system assumes that is has understood and informs the user about which values were filled in ("I understood a specific parameter").

Repetition The system asks the user to repeat his utterance because nothing of what was recognised could be matched with anything in the internal keyword database, or the system reacts to a user command where he explicitly asked the system to repeat its last prompt.

Provide help The system reacts to a request for help on the user side. Help can either be triggered explicitly by the user (e.g. "Which lamps can I operate?") or be provoked implicitly because the user made at least two incomprehensible utterances in succession in certain situations (e.g. he has not managed twice to specify the location of a specific lamp, which is interpreted by the system such that it should support the user by stating the possible options that it can process in the current situation: "The following lamps can be operated: all lamps, the table lamp, the white or the yellow floor lamp.").

No input The system did not receive any input either because the user spoke too low or because he did not utter anything at all for about 20 seconds. In such cases the dialogue manager assumes that the user feels insecure or is not sure about what to say and will consequently react with: "I could not hear you" and subsequently provide some help to the user. Here help fits into the current context of the dialogue.

System non-understanding and out of context The system could not match any of the user input and hence informs the user about its non-understanding by saying: "I could not understand you". This is then followed by a help prompt providing the user with all the options he has at the current point of the interaction.

Dialogue dead end management strategy In case no solution can be found, the system provides alternatives within the current context and asks the user to choose a suitable one.

Confirmation strategy The user verifies the content of a given slot for irreversible actions.

Dialogue termination strategy The system decides when a solution can be proposed and therefore terminates the dialogue.

A strategy to deal with incoherencies The system received contradictory values and enters a clarification dialogue.

Cancelling Cancelling and restarting the dialogue on request by the user.

5. System improvements

After the first tests series with the INSPIRE system, several improvements of the system were implemented to improve the usability of the system. The most important updates are listed below:

Reset patterns: a function which allows a system to restart at any given point in the dialogue. This feature is usually used when the dialogue flow breaks down.

Generic help v. context sensitive help: the system help function was redesigned from *universal* or *generic* to *context sensitive*, i.e. whenever the help function is triggered it provides help regarding the current dialogue context (e.g. in case the interaction focus is set to TV, the contents of the help will only be related to TV-related issues and not present all the possible system functions as was the case with the early system version).

Longer prompts v. shorter prompts: system prompts which users and developers regarded lengthy were successfully shortened resulting in a reduction of the average dialogue length. There are also slight, but noticeable differences regarding the wording of the prompts.

Custom actions: custom actions are personalised sequences of action: a function which allows the operation of various devices by a single command, i.e. the possibility of preparing certain default-settings which will trigger the actions (e.g. switch off all lamps at once).

The changes which were made between the two system versions improved the system notably: as shown in the test persons' behaviour – they were more relaxed; and it also showed in the overall ratings they gave to the system versions. All of these features led to an easier and more relaxed operation of the system.

6. Experiments

6.1. Main objectives of the experiments

The two major test series conducted with the INSPIRE system focused on "metaphors" and "minimal word accuracy".

The metaphors in question are listed below:

Intelligent devices: the user is led to believe that he is talking to single devices. This environment is reached by producing the sound very close to the devices, thereby suggesting that the actual devices "speak" to the user.

Invisible assistant: the metaphor featuring the invisible assistant is referred to as ghost metaphor because of impression of an invisible character whose voice seems to emanate somewhere from offstage. It was established by the activation of several loudspeakers at the ceiling, thus generating a more or less diffuse sound-field.

Talking head: this metaphor includes a visible assistant, a so-called avatar (or: talking head) which is presented as a

face on a computer screen. The corresponding sound output is produced near the screen.

The experiments showed that users preferred the ghost metaphor (Möller et. al., 2004). A large proportion of the users mentioned, however, that they had difficulty in telling the intelligent devices and the ghost metaphor apart. The talking head metaphor was rejected by most. The participants explained that the avatar made them feel forced to do or say something. This made them feel somewhat uncomfortable. It was then decided to equip all subsequent system versions with the ghost metaphor.

The aim of the second set of experiments was to evaluate the influence of the word accuracy level on the users' quality evaluation. Thus, the effects of different word recognition rates were tested, i.e. the main objective was to assess how high the word recognition accuracy should be in order to maintain more or less fluent operability. The result was that "the lowest limit for a positively perceived interaction with the INSPIRE system was 86%" (Krebber, 2005). This was shown by the relevant ratings the users gave to the system. 60% word accuracy is the rate at which the system is still operable (i.e. the dialogue is kept running), although it is not given positive ratings by the users as regards acceptability.

6.2. Test procedure

The tests have been carried out at a living-room test site implemented at Ruhr-University of Bochum and at Philips HomeLab test site. Both test sites provided full control over the majority of experimental factors, such as the system characteristics, the physical test environment, the characteristics of the test subject group, the tasks which had to be carried out by the test subjects, and the rating procedure. All test sites consists of a room decorated in a way which is typical for a living-room (with a couch, armchairs, a low table, shelves) and is equipped with home appliances controllable by means of the current dialogue system: the TV, the VCR, the electronic programme guide (EPG), the answering machine, three switchable lights (two of which are also dimmable) the fan and the blinds. TV, VCR and EPG form one unit. The room is equipped with six loudspeakers for speech prompt output and device feedback. The devices can be controlled from an operator room which is connected to the test room via multi-core cables.

All experiments consisted of four parts:

1. An introduction to the purpose of the experiment, and to the INSPIRE dialogue system and its capabilities
2. An initial questionnaire in which subjects were asked about their personal background, their experience with speech technology in general, and their expectations towards the INSPIRE dialogue system
3. A series of three scenario-guided interactions with the dialogue system; after each interaction, the subject had to fill in a questionnaire with 34 judgments on different quality aspects
4. A final questionnaire by means of which the overall impression of the interactions with the system was collected, the fulfilment of the individual expectations, further suggestions for improvement, etc.

Each of the three interactions carried out by a test subject reflects a specific test condition which corresponds to a certain metaphor in the first test or specific word accuracy (WA) and word error rate (WER) in the latter.

In order to carry out meaningful interactions with the INSPIRE dialogue system, test subjects need to have a specific motivation. In a real-life situation, this motivation arises from the subject's natural behaviour or circumstances of life; in a laboratory test situation, explicit tasks can be given to the test subjects which represent typical actions users are likely to perform. The tasks were defined beforehand, and they are embedded in the three scenarios mentioned above. By means of these scenarios, subjects experience the usage, the functionality and the purpose of the INSPIRE home system.

Test subjects were presented with three scenarios, one for each test condition / recognition rate. Each scenario contained 13-15 tasks which were embedded in a "story", and thus had to be carried out subsequently. The tasks within the three scenarios are very similar, but they appear in different orders to avoid possible learning or fatigue effects. Each scenario addresses at least once the blinds, the lamps, the fan, the TV and the answering machine. In this way, the test subjects have to operate both simple and complex devices in each scenario.

7. The communicative situation

7.1. User expectations

The humans' knowledge about the machine's conversational abilities is established by recalling expectations that have been made before the interaction and from previous experience with systems of similar kinds. While interacting with the dialogue system, users will soon realise that they have to adjust to the system's capabilities in a certain way. This adjustment is based on natural behaviour, i.e. it is learned by every human being who interacts with other human beings with different preconditions on various levels (such as poor hearing or imperfect command of the language).

Seen from this perspective the system is viewed similar to a human conversation partner: this shows in the users' usage of vocabulary, e.g. they find out which keywords are known by the system and will therefore lead to the desired action. This adjustment regarding use of vocabulary takes place very quickly (about 10 to 30 minutes of interaction with the system). When users feel that the system does not understand them acoustically they often start adjusting their pronunciation by hyper articulating. This is a logical consequence of their experiences with humans who have either impaired hearing or who do not have a good command of the interaction partner's language. Every person acting in a communicative way has a general intention which is getting a message across, which means to communicate something to their counterpart.

There are numerous cases where it could be observed that users found their expectations disappointed while interacting with INSPIRE. Disappointed user expectations can lead to frustration and annoyance during the interaction with the system because they receive unexpected answers from the system or they do not understand what the system expects from them. This leads to unwanted prolongation of the dialogue and to their non-fulfilment of wishes. When user expectations are not met, the user is dissatisfied. This leads to new communicative actions on the user's side as he tries to rectify the dialogue process. The system subsequently tries to

identify the user intention. In case it does not succeed (e.g. not enough parameters are recognised or contradictory values are received), the system addresses the user again.

Nevertheless, during the interaction with the machine users adjust their expectations and they undergo a process of learning which equips them with knowledge for further interactions. There has been hardly any research regarding user expectations towards the actual dialogic events within spoken interactions between humans and machines.

However, one publication regarding user expectations was released by Daniel Västfjäll (2003). Västfjäll presents results from experiments which were conducted to evaluate the sound quality of certain products such as machine saws. Even though this experiment does not evaluate the quality of a dialogue system it still tells us a lot about user expectations of auditory events. Västfjäll, notes, for instance, that expectations are always influenced by previous experiences with similar products. The questionnaires from the INSPIRE experiments contain a question which asks for previous experiences with spoken dialogue systems. Considering the users' ratings of the previous system compared to the overall ratings of the INSPIRE system, we can tell that on the whole their expectations, at least regarding overall quality, were met, if not exceeded. Västfjäll lists the following factors which have notable impact on user expectations:

- Product information
- Branding / advertisement
- Visual design
- Knowledge / experience of product
- Feelings towards product / product category
- Tactile / visual information

We can thus conclude that expectations vary according to the individual's history regarding previous experiences and exposure to similar products respectively. In order to achieve maximum user acceptability it is necessary to match the users' expectations towards a SDS as closely as possible. This has proved not always to work out. We can state that the longer it takes the user to reach his aim, the more problems the user has. It seems to be that the more technically advanced systems of this kind become, the more users expect from the system capabilities. It was also observed that those users who had watched many science fiction programmes on TV gave more negative ratings to the system than those who have never consciously seen any dialogue systems at all.

7.2. Dynamic system limitations (linguistic)

References relating to static objects

The most challenging question is definitely that of references of different kinds. As mentioned before, the system cannot resolve references in the same way as humans can. If you look at the system closely, it cannot resolve any references at all since the resolution of references is made possible by a knowledge of the world which humans have established by having made experiences. There are several mechanisms for resolving anaphoric references with pronouns in computational linguistics. Gieselmann (2004) lists various approaches that have been developed between 1977 and 2003. Most of them work with rules which are based on pronouns and their antecedents.

Gieselmann concludes that all of the reference resolution algorithms are reported to resolve more than 90% of all references. She also adds that all those mechanisms were developed based on corpora of written text. For spoken natural language dialogues within human-machine interaction those mechanisms have to be adapted and altered according to differences which exist between spoken and written language (i.e. changing of the grammar structures and covering of spontaneous effects). Additionally, Gieselmann states that there are only very few dialogue systems so far that actually make use of reference resolution because for the last few years, dialogue systems were mainly used for call-routing or information retrieval. For systems that were designed to master more complex tasks such as INSPIRE, reference resolution would be of great use: as already mentioned above, interpretation of temporal, thematic and spatial context reduces confusion since ambiguities will be handled with greater ease and users will not have to repeat themselves as often for reasons of clarification of the dialogue flow. The following extract illustrates references which point to items across several turns:

S1: What else can I do for you?
U1: Switch on the lamps.
S2: I understood lamp as device and switch on as task. Which lamp do you want to operate?
U2: Left.
S3: *switches lamp on*
S4: What else can I do for you?
U3: Reduce the brightness of the lamp.
S5: I understood lamp as device and down as task. Which lamp would you like to operate?
U4: Left again.
S6: *dims lamp*

Table 1: Referring to lamps.

In this sample the user is supposed first to switch on a lamp and adjust its brightness afterwards. He switches on the left floor lamp by specifying its location ("left") (U2). After this has been done he refers to the same lamp by saying "reduce the brightness of the lamp" (U3), but does not specify the exact location. The system, however, has started a new dialogue and therefore deleted the history. Hence it cannot resolve which lamp the user addresses, although for a human this would be no problem at all.

References to dynamic objects

Dynamic objects are objects underlying frequent changes such as movie titles in a TV database. It happens that users refer to the programme they want to choose with proper names (the actual titles) when operating the TV or VCR. Instead of stating the desired parameters (e.g. day and time of broadcast), they might say "I want to see Tagesschau" or "Record Star Trek", while the system will ignore these indications. Humans, on the contrary, can resolve the titles easily. Every adult German, for instance, is certainly aware of the fact that the news programme "Tagesschau" is shown every day at 8 p.m. on the channel ARD. Regarding "Star Trek", things get a little more difficult, but still that programme is known well enough for us to be able to tell that our interaction partner is talking about a movie. So at least the feature movie could be extracted by the system by only being told the title of the programme. The machine does not have any world knowledge as humans do (pool of experience).

Homonyms and ambiguities
Homonyms[1] are a difficult task for a system developer. If interpreted in the wrong task context, the system will fail to interpret the keyword correctly, thereby producing a mismatch and a breakdown of the dialogue flow. While experiments with the system were conducted two frequently occurring homonyms are the words "Nachrichten" and the word "heute". The word "Nachrichten" means news and messages. Therefore, depending on the context, this keyword can trigger the task focus to the answering machine as well as to the TV. Consequently it was frequently the case that the system entered a task focus which was different to that expected by the user. The word ''heute" refers to "today" as well as to the news programme "heute" which can result in a similar confusion.

Ambiguous indications evoke confusion as in: "I would like to operate the lamp next to the TV set". Expressions of this kind confuse the system, since two keywords depicting devices are mentioned by the user: the lamp and the TV set. At its current stage it is impossible for the system to tell that "next to the TV set" is merely an indication of the location of the lamp and does not refer to the TV set as such. In this case the natural language understanding module failed and one of the input words is either matched to a different context or completely ignored. The system therefore lacks the ability to interpret syntactic and semantic context. For humans, on the other hand, hardly any ambiguities exist. The meaning of a word can easily be determined by means of context interpretation.

Synonyms
Synonyms[2] are another issue which occasionally evokes dialogue flow breakdowns, often due to confusion on the user's side. Some users have explicitly stated that they expect that the same words they used should be used by the system, too (as usually happens in interactions between humans). The system, however, lacks the desired flexibility to adjust to the users need in this respect. Please consider the following sample dialogue:

U1: Electronic programme guide.
S1: I understood programme as task. On which day is or was the programme broadcast?
U2: Turn on help function.

Table 2: Confusion by synonyms.

This is a typical example of the system using different words from the user which results in bewilderment on the user's side and causes a breakdown as a consequence of which the user initiates a help request (U2). As a matter of fact the system has processed the user utterance correctly. The user, however, does not comprehend because he expects the system to confirm his utterance using the same words as he does (he would have expected the system to say something like "I understood electronic programme guide").

Complex grammatical constructions
Especially elderly people used grammatical constructions which could not be correctly processed by the system such as the participial construction *"I want the fan turned on"*. This kind of phrasing is not supported by the system at its current stage. The system's interpretation is based on keyword concatenations such as "turn on fan". The same holds for negations as in "not the fan". From a sentence like this the system would only extract the device name and ignore the negation.

7.3. Static system limitations (technical)

The second great category is made up by system limitations of a technical kind in contrast to linguistic or dialogic problems. Of course, the problems to be outlined now do have consequences regarding the dialogue flow, but the reasons are to be found in the system architecture.

Misrecognitions
Four of the 24 participants from the set of experiments asked for improvement of recognition. Since almost 100% recognition was provided in that phase (due to manual transcription), this means that something else evoked the impression that recognition was poor, i.e. there are not enough keywords in the database or the dialogue structure was not transparent enough. However, in the second set of experiments, misrecognitions did happen, due to implementation of the noise generator. Misrecognitions as such are not a problem of the system modules which process language, but are located at the acoustical front-end, namely at the ASR. For instance, a user might tell the system to delete a message from the answering machine, but the system keeps replying that it does not understand which might be due to the user mentioning a phrase which has either been misrecognised by the ASR, lacks known keywords or cannot be interpreted by the system for other reasons. The following extract illustrates a typical case (here misrecognitions by the ASR create confusion):

S1: I understood down. Please state the device you would like to operate.
U1: a sleepless.
S2: I could not understand you. Your options are: the fan, the blinds, the lamps, the TV set or the answering machine.

Table 3: ASR misrecognition.

The misrecognition occurred in U1 and brought about a word which is no known keyword as a consequence of which the system initiates a repair turn. A dialogue like this results in annoyance or even frustration on the user's side if it occurs too often within one interaction (especially problematic when the word "help" is not recognised).

Modality
Although users are made aware of the fact that the system reacts to one input modality only, namely speech, at times they still find themselves disappointed because the system

[1] A homonym is one of two or more words which are spelled and pronounced alike but depict different concepts (as "well" meaning both the adverbial inflection of good and a waterhole.
[2] Synonyms are two or more different words which designate one and the same object.

does not react to physical pointing at things or vocal deictic expressions (e.g. “this lamp here on my right side”). Furthermore a large share of the participants stated that they would like to have an additional remote control at their disposal.

Excessive demand

Consider the following, rather frequently occurring situation: the user is at a point in the dialogue where he is supposed to state a number from a list displaying optional movies for tonight. Some subjects tried to include an additional command in that situation which resulted in an input phrase of this form: “More information on three”. The user tries to include two commands in one utterance: firstly, he picks a movie from a list (three) and, secondly, he indicates that he would like to view detailed information regarding the movie number three. Whenever users are to state a number from a list of options, the system is constrained to react to numbers only, which is due to the focus set. At many points the users are probably not even aware of the fact that they are uttering more than one command in one input utterance or they cannot differentiate when they can give several keywords and when they cannot. If only focused on a single task the system is able to process different values at once, but if the user wishes to fulfil several tasks at once, the system cannot process that input. Even though users probably find out what the inabilities of the system are after a while, they forget when they are concentrated or excited or too involved in the interaction process. This user behaviour is to be attributed to a lack of awareness regarding the system’s limited abilities (i.e. restricted keyword recognition within a certain focus). Being able to operate several devices at a time is still a feature that many users strongly demand.

8. Conclusion

Two researchers of the INSPIRE project discuss their experience and knowledge they gained together with other researchers from work with the INSPIRE system. The focus is on the communicative situation between user and SDS and the gap between user expectation and system realization. Beside of the critics within this article it shall be stated that the INSPIRE system was a success in means of gaining knowledge about SDS and SDS evaluation The successful project will be continued in the future.

Acknowledgment

Special thanks to Paula Smeele (TNO), who died 8th of March 2005 after severe illness. She was a cooperative and productive partner even after the project had ended. The presented work has been carried out in the framework of the EC-funded IST project INSPIRE (IST-2001-32746). The authors would like to thank Alex Trutnev (EPFL), Mirek Melichar (EPFL), Dietmar Schuchardt (ABS), and Hardy Beasekow (ABS) for the technical support, and Heleen Boland (Philips) and Anders Krosch (IKA) for their support in the experiments. The experiments were carried out at the IKA (Prof. J. Blauert, Prof. R. Martin, Prof. U. Jekosch and PD S. Möller) and at the HomeLab at Philips (Jettie Hoonhout).

References

[1] Bernsen, N.O., Dybkjær, H., Dybkjær, L. (1998). Designing interactive speech systems: From first ideas to user testing, Springer, D-Berlin.

[2] Fraser, N.M., Gilbert, G.N. (1991). Simulating speech systems, Computer Speech and Language 5, pp. 81-99.

[3] Gieselmann, P. (2004). Reference Resolution Mechanisms in Dialogue Management. In Proceedings of the CATALOG Workshop, Barcelona, Spain.

[4] Jekosch, U. (2000). Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Habilitation thesis, Ruhr-University, Bochum, Germany.

[5] Krebber, J. (2005). “Hello – Is Anybody at Home?” – About the minimum word accuracy of a smart home spoken dialogue system. In Proc. EUROSPEECH ’05, Lisboa, pp. 2693 – 2696. Lisboa, Portugal

[6] Möller, S. (2003). Quality of telephone-based spoken dialogue systems. Springer, NY-New York.

[7] Möller, S., Krebber, J., Smeele, P. (2004). Evaluating System Metaphors via the Speech Output of a Smart Home System. In 8th Int. Conf. on Spoken Language Processing (Interspeech 2004 - ICSLP), Jeju Island, Korea.

[8] Rajman, M., Bui, T. H., und Portabella, D. (2004). Automated Generation of Finalized Dialogue Based Interfaces. In Swiss Computer Science Conference SCSC04: Multimodal Technologies, Bern, Switzerland.

[9] Trutnev, A., Rajman, M. (2004). Comparative evaluations in the domain of automatic speech recognition. In Proc. 9th International Conference on Language Resources and Evaluation, Volume 4, pp. 1521-1524. Lisboa, Portugal.

[10] Trutnev, A., Ronzenknop, A., Rajman, M. (2004). Speech recognition simulation and its application for Wizard-of-Oz experiments. In Proc. 9th International Conference on Language Resources and Evaluation, Volume 2, pp. 611-614. Lisboa, Portugal.

[11] Västfjäll, D. (2003). Tapping into the personal experience of quality: Expectation-based Sound Quality evaluation. In 1st ISCA Tutorial and Research Workshop on Auditory Quality of Systems, pp. 24–28. Herne, Germany.