Linguistic Adaptations in Spoken Human–Computer Dialogues

Empirical Studies of User Behavior

Linda Bell

Stockholm 2003

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av filosofie doktorsexamen fredagen den 3 oktober kl.11 i sal D3, Lindstedtsvägen 5, KTH, Stockholm.

TRITA-TMH 2003:11
ISSN 1104-5787
ISBN 91-7283-579-6

Cover photo and design by Joakim Gustafson

Abstract

This thesis addresses the question of how speakers adapt their language when they interact with a spoken dialogue system. In human–human dialogue, people continuously adapt to their conversational partners at different levels. When interacting with computers, speakers also to some extent adapt their language to meet (what they believe to be) the constraints of the dialogue system. Furthermore, if a problem occurs in the human–computer dialogue, patterns of linguistic adaptation are often accentuated.

In this thesis, we used an empirical approach in which a series of corpora of human–computer interaction were collected and analyzed. The systems used for data collection included both fully functional stand-alone systems in public settings, and simulated systems in controlled laboratory environments. All of the systems featured animated talking agents, and encouraged users to interact using unrestricted spontaneous language. Linguistic adaptation in the corpora was examined at the phonetic, prosodic, lexical, syntactic and pragmatic levels.

Knowledge about users' linguistic adaptations can be useful in the development of spoken dialogue systems. If we are able to adequately describe their patterns of occurrence (at the different linguistic levels at which they occur), we will be able to build more precise user models, thus improving system performance. Our knowledge of linguistic adaptations can be useful in at least two ways: first, it has been shown that linguistic adaptations can be used to identify (and subsequently repair) errors in human–computer dialogue. Second, we can try to subtly influence users to behave in a certain way, for instance by implicitly encouraging a speaking style that improves speech recognition performance.

"When *I* use a word," Humpty Dumpty said, in a rather scornful tone, "it means just what I choose it to mean – neither more nor less."

"The question is," said Alice, "whether you can make words mean so many different things."

"The question is," said Humpty Dumpty, "which is to be master – that's all."

Alice was much too puzzled to say anything; so after a minute Humpty Dumpty began again:

"They've a temper, some of them – particularly verbs, they're the proudest – adjectives you can do anything with, but not verbs […]"

(from Lewis Carroll's *Through the looking glass*, 1871)

Table of Contents

Acknowledgments

I want to express my gratitude to a number of people, who in different ways have contributed to the writing of this thesis. Thank you all!

First of all, I want to thank my KTH supervisors Rolf Carlson and Björn Granström. Special thanks to Rolf for encouraging me to explore the field of spoken dialogue, allowing me to pursue my research interests, and guiding me through the final stages of the thesis process. Without Rolf's support, this thesis would never have been written.

For constructive criticism and useful discussions following my licentiate seminar, I am thankful to Niels Ole Bernsen at Odense University.

This thesis is based on work done in collaboration with Joakim Gustafson. Our joint research efforts in various projects at KTH and Telia have been challenging, inspiring and rewarding. During moments of doubt, Joakim's curiosity, enthusiasm and seemingly unlimited energy helped me remember what research is all about – and why it is so much fun. I am also grateful for Joakim's help in writing this thesis, where he provided me with thousands of interesting articles, dug deep into old data, suggested various improvements on several chapters and finally struggled hard with the formatting and layout of the text.

I want to express my deep gratitude to Nikolaj Lindberg, whose guidance in the process of writing this thesis has been invaluable. Nikolaj spent countless hours helping me elaborate on the content and form of the thesis, and his meticulous readings of numerous draft versions have resulted in substantial improvements, at all levels. Most of all, I wish to thank Nikolaj for believing in this project, and for managing to convince me I could go through with it.

I am grateful to all the volunteer subjects who contributed to this thesis by lending us their time, and providing us with data. Thanks also to the hard-working dialogue annotators Ludwig Bell, Erik Carlsson and Anna Decker.

My former colleagues at KTH and the Department of Speech, Music and Hearing provided a friendly and helpful work atmosphere. Special thanks to the members of the AdApt group and the spoken dialogue systems seminar. Thanks also to the helpful administrative staff, Cathrin Dunger, Caroline Bergling, Armid Peltomaa, Ebi Rohani-Makvandi and Markku Haapakorpi.

Thanks to Leif Bengtsson, Johan Andersson and Anders Lindström at Telia Research (now TeliaSonera) for giving me the opportunity to work in a stimulating research environment, and for supporting me in the finishing stages of this thesis work. Special thanks to Anders, whose enthusiasm and unyielding faith in this project have been invaluable. Thanks also to my colleagues in the 'Voice Technologies' group - Johan Boye, Robert Eklund, Joakim Gustafson, Anders Lindström, Bertil Lyberg and Mats Wirén, as well as former members Christel Ekvall and Martin Eineborg. In addition, I want to thank Elenore Nordlund and Juliana Handberg at TeliaSonera.

A number of people have helped me by commenting on earlier drafts of this thesis. Many thanks to Johan Boye, Robert Eklund, Lena Enblom, Mattias Heldner, Anders Lindström, Beáta Megyesi, Gunnel Melchers and Mats Wirén. Special thanks to Gunnel Melchers and Lena Enblom for their careful readings of the entire manuscript. Thanks also to Mattias Heldner for helping out with proofreading and practical assistance in the final stages of this work.

For friendship and moral support, I am grateful to Christer, Maria, David, Lena, Mattias, Johanna, Åsa, David, Gastón, María, Noomi, Olle, Elisabet, Laura, Klara, Sara, Hanna and Magdalena.

Finally, I want to thank my dear family. I am grateful to my parents, Ann and Bob, for their constant love and support, and for always encouraging me to read. Thanks also to my big brother Max and his family, and my little brother Ludwig and his rock band. Special thanks to my sweet sister Julia for her generous help (including babysitting at crucial moments). Many thanks to the American family – my grandmother Budd, faster Bonnie and godmother Linda with families. Most of all, I want to thank my husband Johan for his everlasting patience and loving support, and our wonderful daughter Olga for being the sunshine of our lives!

Part of the research presented in this thesis has been carried out at the Centre for Speech Technology (CTT), supported by VINNOVA (the Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. Support has also been provided by the EU/HLT funded project NICE (IST-2001-35293).

Part I

Chapter 1

Introduction

Spoken dialogue systems and multimodal systems with speech interfaces are currently being developed for a broad range of domains and applications. Ideally, such systems allow their users to interact using unconstrained natural language. The long-term goal of research in speech technology is to create interfaces that allow speakers to say anything they want, in any way they want. The perfect speech enabled system should require no prior knowledge or training, thereby making human–computer interaction effortless and efficient.

Early examples of applications for spoken language technology include different types of technical aids for the disabled. For instance, text-to-speech synthesis and a screen reader program make it possible for people with visual impairments to get newspapers, textbooks or websites read out loud. Synthetic speech can also be used as an artificial voice for people who cannot speak. Furthermore, voice command systems for the motorically disabled have been developed for several applications, including text editing and engineering. Another field in which there was an early demand for speech technology was in challenging or potentially dangerous workplaces. One advantage of using speech as an interface is that it can reduce the cognitive load of the user, allowing him or her to focus on the task at hand. This can be crucial in certain demanding environments, such as micro surgery or aeronautics.

More recently, developments in speech technology have enabled users from the general public to use spoken and multimodal interfaces in a variety of situations. Non-experts, computer novices and elderly people are target user groups for the current and coming generation of spoken dialogue systems, thus little or no training should be required to use them. Today, speech technology is also integrated into a number of common services and products. In many cases, timetables for trains and weather forecasts can be quickly (and sometimes only) accessed through simple dialogue systems. For certain languages, software for word processing is delivered with an integrated speech interface, enabling users to switch to voice commands or dictation.

Spoken dialogue interfaces will increasingly become part of our everyday life. Regardless of whether we are users with special needs, professionals or merely in need of timetable information at an odd hour, we are bound to encounter speech interfaces, as they will be part of many of the communicative interfaces of the future. However, the capabilities of today's systems do not always match user expectations. Automatic speech recognition is an error prone technology and the fact that many of the commercial applications are designed to be mobile and used in noisy environments, worsen this problem.

Much research remains to be done in the development of speech interfaces that are intuitive and easy to use, also under difficult conditions. In particular, it is important to increase the understanding of user behavior and interactional strategies in speech based systems. Conveying the system's limitations to a user is a difficult task, and dialogue system designers have often focused more on the technically challenging aspects of the interface. When the human–computer interaction becomes problematic or breaks down, users can become frustrated and even terminate the dialogue.

As people engage in spoken dialogue with a computer system, their general behavior and language strategies are often modified along the way. Either speakers adapt without being aware of it, or they deliberately adapt their language in a manner they believe will make their human–computer dialogue more successful. Both types of adaptations affect users' speech on several levels, involving phonetic, prosodic, syntactic, lexical and pragmatic aspects. For instance, many speakers make an effort to speak clearly and avoid complex grammatical structures when they interact with computers. Such deliberate linguistic adaptations can be understood as the user's way of assessing the capabilities of the system. Depending on how the human–computer interaction evolves, these adaptations may then become more or less accentuated throughout the course of the dialogue. If everything runs smoothly, the user's language is likely to become less computer-directed in style. On the other hand, system errors often elicit unintentional or deliberate modifications of user behavior. Apart from leading to breakdowns in human-machine communication, frustration and anger, system errors and resulting spirals of problematic interaction often magnify patterns of linguistic adaptation.

As speech technology advances and spoken dialogue systems become more sophisticated, people's expectations of these systems' capabilities will increase correspondingly. When a spoken dialogue fails to meet these expectations, and a problem occurs, users attempt different strategies to recover from this situation. These user strategies often involve altering or adapting one or several aspects of the spoken language input to the system. Since user expectations are likely to always be as high as or even higher than the capabilities of each currently developed system, the question of how these adaptations manifest themselves becomes an important one. If, as will be argued in this thesis, these adaptations are foreseeable and predictable, it should be possible to integrate a method of handling them into future systems.

1.1 The Challenge

In this thesis, we address the question of how people adapt their language when they interact with a spoken dialogue system. Spoken natural language as a means of communicating with a computer is often advocated because it is natural and effortless, especially for inexperienced users. A complicating factor is that during human–computer interaction, speakers adapt to (what they believe to be) the linguistic capabilities of the system. In addition, they fall back on human discourse strategies, especially if they encounter some problem in their dialogue. Without knowledge about users' linguistic adaptations built into the systems, they will never attain the level of naturalness which they were aimed for. The next generation of spoken dialogue systems must include knowledge both of how users adapt in today's spoken dialogue systems, and knowledge about how they adapt in accordance with the structures of human–human dialogue.

Linguistic adaptations are a natural and to some extent inevitable part of human communicative behavior. All spoken dialogue systems contain some model of the user's behavior, often as separate models for different linguistic levels. On the one hand, linguistic adaptations can be regarded as a problem, because they contribute to the variability of user behavior. On the other hand, they can be useful in the development of spoken dialogue systems. If we are able to adequately describe their patterns of occurrence (at the different linguistic levels at which they occur), that knowledge can be used to build more precise user models, thus improving system performance. Furthermore, our knowledge of linguistic adaptations can be useful in at least two ways: First, it has been shown that linguistic adaptations can be used to identify (and subsequently repair) errors in human–computer dialogue. Second, we can try to subtly influence users to behave in a certain way, for instance by implicitly encouraging a speaking style that improves speech recognition performance.

Although there are recurrent patterns in how linguistic adaptations are manifested in spoken human–computer dialogue, it is also clear that individual differences are great. Just as in human discourse, speaking styles and dialogue strategies vary from one user to another. For example, certain users are likely to voluntarily give a spoken dialogue system feedback throughout the dialogue, while others have to be explicitly asked to provide the same information. Current systems are limited in their capacity to predict and handle different user strategies. Future systems will hopefully become better at modelling these individual differences in linguistic adaptations.

1.2 Thesis Contributions

This thesis addresses the question of how speakers adapt their language during spontaneous spoken human–computer dialogues. The approach has been empirical: a number of computer-directed speech corpora were collected, labeled and analyzed. The dialogue corpora examined in the thesis include data collected with simulated or semi-simulated systems in controlled settings as well as data collected with fully functional systems in public environments. All of the systems used for data collection had advanced spoken dialogue interfaces, in which users were encouraged to interact spontaneously using unrestricted spoken language. Each of the systems had an animated agent with whom the user interacted face-to-face. Although several other aspects of the data were also studied, the goal of the empirical studies reported in this thesis has been to examine speakers' linguistic adaptations in the dialogues. More specifically, these have been the findings of the present study:

- Results for spontaneous spoken Swedish human–computer interaction support previously reported findings:
  - During error resolution, users' speech moves toward hyperarticulation, is increased in duration, contains a greater number of inserted pauses and is more careful in pronunciation
  - Users' choices of lexical items and syntactical structures mirror the language used by the system
- During spoken interaction with animated agents, the following patterns in users' linguistic adaptations were also observed:
  - Users adapt their speech rate to that of the system's output, in particular if the system speaks slower than normal
  - A spoken dialogue system that uses and responds to greetings and social remarks will also encourage its users to engage in socializing
- Adults and children exhibit partly different patterns of adaptive behavior:
  - Children seem to have access to a limited repertoire of linguistic strategies, and during error handling resort to repeating the same utterance over and over again. Adults, on the other hand, try various strategies during their interaction
  - Children often speak louder or even scream when the system fails to understand them

The present thesis examines user behavior and linguistic adaptations in spoken and multimodal dialogue systems. Although adaptations may (and do) occur in other input modalities as well, those that do not manifest themselves in spoken language will be disregarded in the present context. When a speaker's use of another modality is likely to have affected the outcome of the dialogue and the type and magnitude of the spoken language adaptation, this has been taken into account. However, this study focuses on the spoken language input of the multimodal systems under discussion.

The corpora of spontaneous spoken language collected and examined in the course of this research are all in Swedish. A few examples of users unsuccessfully trying to address the dialogue systems in other languages can be found in the corpora, but these utterances were excluded from the analyses. Although it is possible that the findings of this thesis are relevant for other, similar languages, the present study focuses exclusively on language phenomena in Swedish corpora.

The spoken and multimodal dialogue systems used to perform the user studies described in this thesis were developed jointly by a number of past and present colleagues at KTH–CTT and Telia Research (now TeliaSonera). In Part II below, this collaborative work is described in detail. The user studies, the results of which are presented in the second part of this thesis, are also the outcome of collaboration with others. Most of the findings of these user studies have been previously published. At the beginning of each chapter describing these experiments, the original source of the results is noted. A complete list of publications by the present author can be found at the end of the thesis.

Before moving on, the reader should be cautioned: this thesis contains no systematic comparison of human–human and human–computer dialogues using identical settings and tasks. Instead, theories of human–human dialogue are used as a sort of backdrop for extensive empirical studies of users' linguistic behavior in human–computer interaction. At a high level, the thesis contains some comparisons between human conversation and computer-directed speech. However, in the sense that no controlled same-domain experiments were performed, this thesis will not satisfy the reader who is looking for the answer to the question of how people interact with other people vs. computer systems.

1.3 Outline

This thesis is divided into three parts, each consisting of one or several chapters:

Part I contains an overview and discussion of related work in the areas of human–human and human–computer interaction, as well a chapter on data collection methodology.

Chapter 1 is an introduction to the thesis.

Chapter 2 introduces speech interfaces, and presents the case for and against them. A brief outline of the research dialogue systems described in this thesis is presented, followed by a few examples of linguistic adaptation from the spoken dialogue corpora.

Chapter 3 considers theories of spoken human–human dialogue that are relevant in the present context.

Chapter 4 is an overview of some previous work on linguistic adaptation in human–human and human–computer interaction. Adaptive strategies at different linguistic levels are discussed, and comparisons between speakers' behavior towards other humans and spoken dialogue interfaces are made.

Chapter 5 continues the review of previous studies. In this chapter, we address the question of how errors which occur in spoken dialogue systems result in user adaptations at different levels.

Chapter 6 contains a discussion about empirical methods and some methodological issues that are relevant for the present study. We make a comparison of different data collection methods and their respective benefits. The chapter also contains a discussion of ethics in the context of data collection with spoken dialogue systems.

Part II consists of a description of four spoken dialogue corpora and a series of investigations, the results of which make up the contributions of the present thesis.

Chapter 7 describes four spoken dialogue systems and the corpora collected with them. Each speech corpus is illustrated by means of a dialogue example. In the last part of the chapter, there is a brief overview of the included systems and corpora, as well as a discussion about some of the relevant aspects of data collection.

Chapter 8 discusses the occurrence of phonetic and prosodic adaptations in three of the included corpora. The first and second studies both contain analyses of repetitive sequences in human–computer dialogue. In these studies, the differences in user strategies between adults and children are discussed in some detail. In the third and final study, we present an empirical investigation aimed at determining whether users' speech rate was affected by the output of the spoken dialogue system they were interacting with.

Chapter 9 consists of two empirical studies in which lexical and syntactic adaptations are examined. The first study examines the general complexity of a corpus of human–computer interaction, and also investigates lexical and syntactic adaptations in non-identical repetitions. In the second study, we focus on how users modify their lexical choices to match that of the system they are interacting with.

Chapter 10 addresses the question of how speakers behave towards computers with respect to social interaction, turn-taking and politeness. Four corpus studies are described. In the study, the implications of users' socializing behavior in human–computer interaction are discussed. In the second study, we describe fragmented user utterances. The third and fourth studies deal with disfluencies and feedback, respectively.

Part III contains a summary, some concluding remarks, and future work.

Chapter 11 summarizes the results from the empirical studies presented in Part II, discusses some implications of these findings and outlines possible directions for future work.

Chapter 2

Speech Interfaces

Spoken dialogue systems are developed to provide a means for people to interact with computers using spontaneous speech. Put simply, a spoken dialogue system is a system that recognizes natural spoken input, interprets its meaning in the current context, decides which action to perform based on that interpretation and subsequently generates an appropriate answer (e.g. by means of a pre-recorded prompt or text-to-speech synthesis). Figure 2.1 below is an illustration of a simple spoken dialogue system.

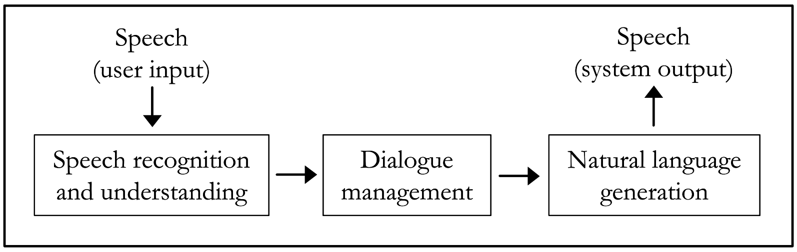


Figure 2.1 A spoken dialogue system.

There are many types of spoken dialogue systems. For commercial applications, single-domain systems with restricted vocabularies and grammars have already been successfully developed. These telephone-accessed systems, which are briefly discussed in the method chapter below, are specifically designed for a narrow task, such as information about bank statements, weather forecasts or time tables (Balentine and Morgan 1999). In order to minimize the risk of errors and increase the system's accuracy, the dialogue structure often restricts the freedom of users' spoken input. For example, a banking system may instruct its users that they can only say 'balance', 'transfer' or 'new account'. In some currently available commercial systems, techniques for key-word spotting (where the system picks out the important words) make it possible to speak more spontaneously. Nonetheless, users are still required to stay within the limits of a well-defined and narrow task.

2.1 Advantages of Speech Interfaces

> "Everyone can communicate by body movements, slightly fewer people can express themselves vocally, still fewer people can communicate by handwriting or drawing, and typing is the least widely available human communicative skill. That order is exactly the reverse of their adaptability to man-computer communication. Typewriting, the least universal mode of person-to-person communication, is easiest and most adaptable for man-computer communication. On the other hand, no-one foresees body movements – the most universal mode of human communication – as a viable alternative for man-computer communication, even in the year 2,001." (Chapanis 1981:69)

Even though most computers are currently operated by means of a keyboard and mouse, this is not the natural order of things. In the more than twenty years that have passed since Chapanis wrote the article quoted above, human gestures and body movements have emerged as an input channel for human–computer interaction. Furthermore, the spread of desktop computers have increased the use of keyboards by several magnitudes. All the same, Chapanis's point remains a valid one. Globally speaking, typewriting remains one of the least widely spread communicative skills and it requires training and learning. In a series of experiments, Chapanis and colleagues explored how people naturally communicate with each other and compared the efficiency of different interaction modes (Chapanis 1981). Not surprisingly, they report that communication-rich mode (face-to-face interaction) is the most effective way of solving a task in as little time as possible. However, when two subjects are separated by a wall and can use voice alone to solve the same problem they are almost as quick and effective. If the goal of a human–computer interface is to allow efficient and quick transfer of information, spoken dialogue is an excellent medium from the point of view of the human interlocutor.

Speech interfaces have a number of advantages. When compared to graphical user interfaces, unconstrained spoken input allows users to immediately pinpoint what they want instead of browsing through numerous menus and clicking on link after link. Ideally, the user is relieved of the burden of trying to figure out how the designer structured the interface. In an early comparison of typed input and speech, Hauptmann and Rudnicky (1990) showed that speech is faster and more efficient as an input modality for strings that require more than a few keystrokes. For motorically disabled people, or for anyone in a hands-busy situation, speech can be an excellent input channel.

When a user needs to be mobile, and it is inconvenient to carry around a screen and keyboard, speech-based interfaces can be very useful. Because of this, and as a result of recent advances in processing power and miniaturization of electronic equipment, interfaces currently developed for mobile applications are often multimodal, offering the user the possibility to tap or click in addition to (or instead of) speaking.

2.2 Limitations of Speech Interfaces

While spontaneous conversation is easy and effortless from the human point of view, it has turned out to be difficult to model artificially, using a computer system. Advances in the field of automatic speech recognition (ASR) made over the past few decades have made it possible to develop systems capable of large-vocabulary user-independent recognition of continuous speech with relatively high performance, even in noisy environments. This is possible using current statistically-based recognizers which have been trained acoustically on very large domain-independent corpora, and then primed on a smaller set of domain-specific data. However, as a long-standing debate in the field shows (Pierce 1969; Levinson 1994; Boulard, Hermansky and Morgan 1996; Greenberg 1998), these systems solve a quite different problem than that of spoken language *understanding*, since they only transcribe the sequence of words given the speech signal. While this may be perfectly adequate for many applications, such as dictation, it normally does not suffice for spoken dialogue systems, where ASR is an input component to higher levels of analysis. In fact, even given a perfect transcription (at the word-level), misunderstandings and problems may occur on several other levels of the human–computer dialogue, levels in which research is even more immature.[1] Moreover, prosodic and emotive features are important carriers of information in human discourse, especially when a problem occurs in the dialogue. Unfortunately, state-of-the-art spoken language technology has yet to utilize these features, partly because these areas have proven problematic to research and model. Therefore, spoken language understanding remains an inherently sensitive and error-prone technology.

[1] The following table exemplifies speech-based systems' understanding capabilities at different linguistic levels. The degree of complexity increases from left to right in the table.

Phonetic/ Prosodic	*Standard pronunciation*	*Pronunciation variations*	*Disfluency understanding*	*Prosody recognition*
Lexicon	*10 words*	*500 words*	*50.000 words*	*500.000 words*
Syntax	*Single word*	*Single phrase*	*Single utterance*	*Multiple utterances*
Pragmatic	*One simple domain*	*One complex domain*	*Multiple simple domains*	*Multiple complex domains, common sense*
Turn-taking	*Directed dialogue, beep*	*Mixed initiative, barge-in*	*Prosodic turn-taking cues*	*Multimodal turn-taking cues*

Speech as a medium for transmitting information is limited by the human capacity for short-term memory, which is estimated to be able to hold material for about 30 seconds (Smith 1993). Miller (1956) showed that the human brain on average can keep track of seven plus/minus two meaningful items. This makes it inappropriate to use spoken language as a medium or storage for larger quantities of condensed information. In addition, when compared to writing, speaking draws heavier on the user's overall cognitive abilities. Accordingly, studies have shown that speech is not optimal for tasks that require real-time planning, such as word processing (Karl, Pettey and Shneiderman 1993; Shneiderman 2000).

It has been questioned whether the goal of natural, unrestricted spoken dialogue with computers is a realistic or even ideal one. According to Karat et al. (1999), speech should not be regarded as 'natural' for communicating with machines. When we interact with our fellow humans, there is a lot of shared context. Humans and computers do not share a great deal of knowledge (yet), something that makes more advanced interaction difficult. The authors also point to the disadvantages of utilizing speech in a user interface, such as sensitivity to background noise and the risk of interference with other parallel cognitive activities. Furthermore, Karat et al. emphasize the effort required in designing computer systems which use speech as their input and/or output and observe that:

> " […] although having no difference between human–human and human–computer communication might be a good goal, it is not one likely to be attainable in the near future." (Karat et al. 1999:2)

Other authors have suggested that it is simply wrong to strive to model human–computer interaction on human–human interaction. Shneiderman (2000) states that there is a limit to the types of conversation people will want to have with computers. For instance, other humans will always be better than machines at giving us emotional support and guidance. In that sense, it may be inappropriate to model human–computer dialogue on human–human dialogue.

Jönsson and Dahlbäck (1988) question the position that a natural language interface to a computer should attempt to resemble a human dialogue. The authors point to a number of differences between human–human and human–computer dialogues that will affect the way people speak to computers. For instance, the social context of a human–computer dialogue does not usually require a speaker to observe human codes of conduct and be well-mannered. If the language used to interact with a computer is different from the language used in dialogues between humans, natural language interfaces should be designed to reflect this fact. Jönsson and Dahlbäcks' (1988) main point is methodological: human–computer interaction is best studied by examining (simulated) dialogues between people and computers, rather than by attempting to mimic human conversation. The advantages and disadvantages of different data collection methods are discussed in Chapter 6 below.

2.3 The Dialogue Systems in this Thesis

The spoken dialogue systems that are described in the current thesis are all experimental research systems, developed with the aim to advance the state of the art by encouraging users to speak spontaneously using their own language. Users of these systems were never explicitly told what to say, or how to say it. Although individually different, the four spoken dialogue systems of this thesis can all be characterized in the following way:

- the spoken dialogue systems were embodied by animated talking agents
- users engaged in face-to-face interaction with the spoken dialogue systems
- the spoken dialogue systems could either handle or simulate the handling of certain social user behavior
- to some degree, the spoken dialogue systems were designed to appear more intelligent than they really were
- users were encouraged to engage in spontaneous conversation with the spoken dialogue systems, and test the limits of the systems' capabilities

The current thesis deals with users' spoken input to spoken dialogue systems. At the technical level, the systems will be treated as 'black boxes' – the internal workings of the speech recognizer or dialogue manager will be discussed only when it is believed to be relevant for how users react to these aspects.

The question of which factors affect users' behavior during human–computer interaction is a complex one, and it is beyond the scope of this thesis to fully cover it. It will be suggested that aspects such as the task, setting, interface design and initial success of the dialogue are contributing factors. One interesting issue is to what degree the users' *beliefs* about the limitations of spoken dialogue systems influence their way of interacting. This question will be further addressed in Chapter 3, which discusses various factors that contribute to linguistic adaptations in spoken dialogue systems.

In Chapter 7, each of the spoken dialogue systems are described in some detail. Furthermore, the spoken language corpora and some aspects of the user interfaces of the systems that may have contributed to the outcome of the dialogues are discussed. However, since examples taken from the corpora are discussed throughout the first part of the thesis, we will now briefly introduce the four systems:

The **August** system was a fully functional spoken dialogue system with an animated agent which was displayed in a public location in downtown Stockholm. The users of the system were given no explicit instructions, but were encouraged by the animated agent August to discuss topics such as speech technology, facts about the city of Stockholm and turn-of-the-century literature. The domain of the **AdApt** system was the real-estate market of Stockholm, with the animated agent Urban functioning as a broker. A simulated version of the system was used to collect data in a controlled laboratory environment. This data was subsequently used for the development and implementation of a fully functional system. The **Pixie** system is part of an ongoing exhibition at the Telecom museum in Stockholm. In a full-scale model of an apartment from a not-so-distant future, visitors interact with the animated agent Pixie who provides access to information and available services. The system is simple but fully functional, and encourages users of all ages to engage in spoken interaction in several domains. **NICE**, a system currently under development, will enable users to interact with different characters in a fairy-tale world inspired by the Danish author H C Andersen. In the first simulated version of the system, visitors at the Telecom museum interacted with the animated agent Cloddy Hans in a simple problem-solving scenario.

2.4 Dialogue Examples

Users often approach spoken dialogue systems with high expectations about these systems' ability to understand and respond to unrestricted language input. To increase our knowledge about what users expect from coming generations of such systems, researchers sometimes let them appear more knowledgeable than they really are. In the development of the August system, which is described in Chapter 7, this approach was put to the test using an unsupervised system in a public location. By sometimes giving general answers to user questions and through a credible animated face with lip-synchronized synthetic speech, August conveyed the impression of being a confused but not unintelligent character. The resulting dialogues, although sometimes disappointing from the user's point of view, revealed a great deal about people's preconceived opinions of dialogue systems and optimistic hopes for the future of speech technology.

To illustrate this point we will now turn to a few examples of linguistic adaptation, taken from the August corpus. Throughout this thesis, corpus examples are often chosen to illustrate how users react to difficulties and problems in the dialogue. These examples are not always representative of the corpora in their entirety. The transcribed excerpts in the tables below are intended to exemplify how high initial user expectations sometimes result in the linguistic phenomena which are the focus of the current study.

Table 2.1 Parts of a dialogue between August and a female user

User	**August**
Hur många spanjorer finns i Stockholm? *How many Spaniards are there in Stockholm?*	Jag hittade inget i min databas *I could not find anything in my database*
Hur många spanjorer eh bor i Sverige? *How many Spaniards eh live in Sweden?*	Jag hittade inget i min databas *I could not find anything in my database*
Vad heter Spaniens huvudstad? /silence/ det **måste** han kunna *What is the capital of Spain? /silence/ he **has** to know this one*	Jag hittade inget i min databas *I could not find anything in my database*

In the excerpt from the August corpus in Table 2.1 above, a female user tries to get some answers from the animated agent. She begins by asking an encyclopedic type of question, the content of which is gradually simplified. When the first utterance is compared to the second, two lexical items have been exchanged. An inserted filled pause in the second utterance shows the uncertainty of the user. In the user's third and final attempt, a related but simpler question is asked. From her last remark, which is not computer-directed but addressed to a friend in the room, it becomes obvious that this user's expectations of this system's capabilities were rather high, and that August disappointed her. From the system's point of view, however, the responses were correct – the database did not contain what the user asked about. The system did indicate to its users that it had some information about Stockholm, but there was nothing there even remotely related to Spaniards or Spain.

In the example in Table 2.2 below, a male user is more successful in getting August to understand him. However, the dialogue contains quite a few misunderstandings and the second through fourth dialogue turn exemplify instances of linguistic adaptation. The user modifies his original question ("Are you alive now") twice, and finally gets some sort of answer from the system.

Table 2.2 A dialogue between August and a male user

User	August
Vad heter du? *What is your name?*	Strindberg var en stor författare vid sekelskiftet *Strindberg was a great author at the turn of the century*
Lever du nu? *Are you alive now?*	Det där förstod jag inte *I did not understand that*
Är du levande? *Are you living?*	Vad är det du vill veta egentligen? *What is it really that you want to know?*
Är du död? *Are you dead?*	Nej då, jag skulle ändå vakna nu *Oh no, I was waking up anyway*
Vad känner du nu? *What do you feel now?*	Hej, jag kommer från institutionen för tal musik och hörsel *Hello, I come from the department of speech, music and hearing*
Var bor du? *Where do you live?*	Du måste ange en gata också *You must name a street as well*
Hur gammal är du? *How old are you?*	Jag fyller hundrafemti år. Har du med dig någon present? *It is my onehundred and fiftieth birthday. Did you bring a present?*
Nej *No*	Det där förstod jag inte *I did not understand that*
Förstår du vad nej betyder? *Do you understand what no means?*	Min bror heter Alf och bor på KTH *My brother is called Alf and lives at KTH*

These dialogue examples conclude the second chapter of this thesis. In this chapter, we have discussed some of the advantages of using spoken dialogue systems, some limitations of speech interfaces, and briefly introduced the systems that are the focus of this study. Finally, a few examples of linguistic adaptation from one of the spoken dialogue corpora were presented and discussed.

In the following chapter, we examine some characteristic features of spontaneous human–human discourse relevant for understanding linguistic adaptation as it is manifested in human–human and human–computer dialogue.

Chapter 3

The Structure of Spoken Discourse

This chapter reviews some of the most important theories describing the structures and regularities of human dialogue and discourse. This overview is relevant as a background for the subsequent chapters which address human–computer dialogue phenomena. After a brief introduction about the fundamental nature of spoken dialogue, the second section deals with some characteristic features of spontaneous spoken language. Next, we consider collaboration in language and how the theory of *grounding* explains language use. Subsequently, we review *turn-taking* phenomena in spontaneous discourse and then focus on *speech act theory* and its implications for analyzing dialogue data. Finally, *conversational implicature* and politeness phenomena are addressed.

3.1 Face-to-Face Conversation

While human–computer interaction is a relatively new form of communication, spoken dialogue itself is not. On the contrary, it is believed that spoken language can be dated back to the birth of humanity itself (Allwood 1996). It has been argued that spoken interaction is the most fundamental type of language. In a comparison between writing and speech, Bloomfield (1933) argues that "[…] writing is not language, but merely a way of recording language by visible marks." Fillmore (1981) states that face-to-face conversation is the primary use of language, and that all other uses of language can be understood through their deviation from that base. As observed by Clark (1996), one aspect of face-to-face conversation that makes it different from all other types of language is that it is common to all known human cultures. Halliday (1989) points out that for 99.5% of the history of the human race, sound was the only medium of expression for language.

Conversational settings do not require speakers and listeners to have special skills apart from those that are part of the process of first-language acquisition. On the contrary, as noted by Clark (1996), it takes years to learn how to read and write and many people never acquire these skills. It is estimated that about a sixth of the world's population live in a language society that does not employ a writing system (Clark 1996). Even in societies where such conventions have been developed, young children, certain groups of disabled people and illiterate members of society communicate by means of spoken interaction alone.

3.2 Spontaneous Spoken Language

The aim of this thesis is to increase our understanding of how speakers who interact with spoken dialogue systems adapt their language when they are addressing a computer. In this context, it is important to keep some characteristic features of spontaneous spoken language in mind.

Despite much interest in language and communication, spoken language phenomena have so far been relatively unexplored. Historically, this can be explained by the fact that written language sources were the only ones available. Nonetheless, even after recording devices were invented, the focus of most studies in linguistics has remained in the area of written language. Part of the reason for this is that the study of speech requires much effort in the areas of collecting, processing and analyzing data. Leech (1991) notes that transcribing speech data is a difficult and time-consuming process, and that collecting speech corpora the equal size of today's written corpora remains a dream of the future.

Yet another reason for the focus on written language was that studies of textual sources and corpora were considered high-status, and given priority. Spoken language was seen as unstructured, superficial, simple and less worthwhile studying (Beattie 1983).

> "Spontaneous speech is unlike written text. It contains many mistakes, sentences are usually brief and indeed the whole fabric of verbal expression is riddled with hesitations and silences."
>
> (Beattie 1983:33, quoted in Halliday 1989:76)

As Biber (1988) points out, many of the apparent differences between text and speech can be attributed to variations of genre. For instance, a formal speech given in the national parliament is likely to contain larger lexical variation than a handwritten note that was quickly jotted down during a lecture.

Spontaneous spoken language is different from written language in many ways. However, it is a misconception to consider written text the norm and spontaneous spoken interaction some deviation from that norm. On the contrary, as mentioned in Section 3.1 above, spoken language is the fundamental and primary form of communication. What is important to consider is that we should not study spoken language phenomena from the viewpoint of patterns and structures found in written text. Spoken language is interesting in its own right and its structures should be independently analyzed.

This thesis deals with linguistic phenomena which are manifested in spontaneous speech. As a general background, it could therefore be useful to get an overview of some of the characterizing features of this type of language. In a cross-linguistic overview of the key properties of spoken language, Miller and Weinert (1998) stress the following features of spontaneous speech:

- it is produced in real-time, and cannot be edited.
- it reflects the limitations of the human short-term memory.
- it is typically produced by people speaking face-to-face.

Since spoken language cannot be edited, it tends to contain hesitations, filled pauses, restarts and other disfluencies. Disfluencies in spoken language should not be seen as mere mistakes or disturbances. Instead, their distribution in spontaneous speech displays regular patterns which are reflections of the speaker's cognitive load, and they can even be carriers of linguistic information themselves. Furthermore, as mentioned in Section 2.2 above, the limitation of listeners' and speakers' short-term memory makes speech inappropriate for tasks that require longer sequences of information to be conveyed in a single turn. Another characteristic of spontaneous speech is that it is produced in conversational settings, with speakers facing one another. When people are not speaking face-to-face, as in telephone conversations, this affects their language production and perception. For instance, turn-taking works less efficiently and disfluency rates are significantly higher in two-person telephone calls when compared to face-to-face dialogue (Oviatt 1995).

Miller and Weinert (1998) go on to list some of the linguistic properties typical of spontaneous spoken language:

- small quantities of information are assigned to each phrase and clause.
- the syntax of spontaneous speech is fragmented, phrases and clausal constructions are less complex.
- the range of vocabulary is smaller than in written language.

(adapted from Miller and Weinert 1998)

Spontaneous spoken language has certain distinguishing qualities, something which is important to keep in mind when designing speech interfaces and analyzing speech corpora. In addition to the properties listed above, Halliday (1989) points to the fact that there are certain linguistic features that exist only in spoken language. These features are known as *prosodic* and *paralinguistic*. While prosodic features are rule-based and part of the linguistic system, paralinguistic features can be described as non-systematic individual variations. For example, intonation and rhythm belong to prosody, and a creaky voice is a paralinguistic feature.

3.3 Language as Collaboration

Clark's theory of *grounding* describes language use as "A joint action […] carried out by an ensemble of people acting in coordination with each other." (Clark 1996:3). Even before two people start talking, they share what is known as a *common ground* (Stalnaker 1978). For instance, the speakers may belong to the same cultural and social community. When one of them says "Ludwig Wittgenstein," she assumes that the other dialogue participant knows she is referring to the legendary twentieth century philosopher and professor at Cambridge. At each subsequent time that these two speakers interact and discuss Cambridge philosophers, their common ground in this area is increased. One of the speakers asserts that although she considers the late Wittgenstein's thoughts on language profound, she finds his early theories less interesting. The other speaker asks what she means by 'theories' – he considers the early Wittgenstein's works compelling from an aesthetic point of view. The first speaker yields, and agrees that the work of the young Wittgenstein is quite interesting. The speakers have thus come to an understanding about what they mean when they discuss the two phases of Wittgenstein's philosophy, and have increased their common ground.

According to Clark (1996), efficient interaction necessarily involves the co-ordination of speaker's meaning and addressee's understanding, so that participants in a conversation continuously work at enlarging their common ground. This may involve the negotiation of what vocabulary, referential expressions, sentence structure etc., to use among themselves.

In a series of studies, Clark and colleagues sought to empirically test and explore the theory of grounding. Clark and Wilkes-Gibbs (1986) showed that speakers often engage in a sort of negotiation as they collaborate on which referring expressions to use. Participants repair, extend or replace the noun phrase under discussion until they are satisfied with the result. In the course of the experiment, pairs of subjects separated by a wall were given identical sets of Chinese Tangram figures which they were asked to describe and position on a chart. Tangram is a Chinese puzzle consisting of a square divided into seven pieces that must be arranged to match particular designs. The Tangram figures are often difficult to describe and the subjects of Clark and Wilkes-Gibbs' experiment, who engaged in conversation to solve this task, worked together and cooperated to minimize their shared effort. In what Clark and Wilkes-Gibbs call the *acceptance process*, the negotiation is carried out in several steps. In this example, person A who originally proposed a way of referring to the figure is not happy with the suggestion and decides to replace it with another:

A Okay, the next one looks, is the one with
the person standing on one leg with the tail

B Okay

A Looks like an ice skater

B Yeah, okay

(from Clark and Wilkes-Gibbs 1986:131)

In subsequent references in the dialogue, the term *ice skater* is used. As the example shows, speakers continuously give each other feedback throughout the dialogue. In this way, they are able to signal understanding and non-understanding, and warn each other of upcoming difficulties. Clark and Wilkes-Gibbs (1986) also note that it is often not necessary to ground every aspect of a communicative process. Instead, participants are satisfied once they reach the *grounding criterion*, which is when the participants in conversation agree they have grounded sufficiently for the current purpose. Depending on the type of communicative task and situation, the grounding criterion will be reached at different points in the interaction.

Clark and Schaefer (1989) describe how people contribute to the discourse they are involved in by adding to the common ground. As they engage in the collaborative process of discourse, participants give each other evidence that they have understood what the other person is saying. Based on empirical studies of the London-Lund corpus of spontaneous spoken English (Svartvik and Quirk 1980), the authors propose a scale of the different types of evidence of understanding that may occur in a dialogue between person A and person B:

- Continued attention
- Initiation of the relevant next contribution
- Acknowledgement (B nods or says 'uh huh', 'yeah' or the like)
- Demonstration (B demonstrates all or parts of what he has understood A to mean)
- Display (B displays verbatim all or parts of A's presentation)

(Clark and Schaefer 1989)

This scale ranges from weakest to strongest. Depending on the task, context and previously established common ground, a contribution in discourse may come in different forms. Participants cannot ground everything, since this would make communication ineffective. On the other hand, if they do not give each other evidence of understanding or non-understanding, their discourse could become problematic later on.

Heeman and Hirst (1995) present a computational model for how participants in conversation collaborate in order to refer to an object in a successful manner. Traum (1994) developed a model of how participants in dialogue come to reach a state of mutual understanding of a speaker's utterance. Traum's computational model makes use of the notions of beliefs, mutual beliefs, intentions and obligations to decide when a dialogue contribution has been properly grounded.

3.4 Turn-taking in Dialogue

When compared to monologue or formalized conversation, spontaneous spoken dialogue between several speakers may appear chaotic and unstructured. Typically, several impatient speakers wait to barge in and interrupt the person currently holding the floor. Turn-taking is often hurried and might seem to be totally irregular. However, empirical studies have shown that what may at first sight seem to be an unorganized blur of voices is in fact a process that is highly predictable and at least to some extent rule-governed.

Even though it may appear as if speakers constantly interrupt one another, overlap seldom occurs in spontaneous speech. Levinson (1983) has shown that people speak at the same time in spontaneous American English during less than 5% of the dialogue. The transition from one interlocutor to another is usually very efficient, involving short or nonexistent pauses. It has been reported that the time between two turns in spontaneous conversation can be as short as a few hundred milliseconds (Bull 1996; Jurafsky and Martin 2000). In an experiment using the Edinburgh map task corpus (Anderson et al. 1991), Bull (1996) reports that about one fourth of the between-speaker intervals were less than 100 ms. This amount of time is as short as or even shorter than the minimum time required for the actual planning of an utterance. It thus appears as if participants in dialogue begin planning their next utterance even *before* the present speaker has finished his turn.

The subtle and fine-tuned process of turn-taking is governed by a set of rules. Sacks, Schegloff and Jefferson (1974) analyzed audio recordings of spontaneous dialogue to understand how conversation is organized. They discovered rules that operate on a turn-to-turn basis, and observed that these rules can be regarded as a method for sharing a scarce resource – the control over the 'floor' (Levinson 1983). According to Sacks et al. (1974), conversation is composed of 'turn-constructional units.' These units are phrases or clauses that are prosodically marked, giving the interlocutors an idea of where in the dialogue it would be possible to barge in and take the turn. There are certain places in the dialogue where it is possible for a new speaker to take over from the one presently holding the floor. They are called *transition-relevance places* (TRPs). Non-verbal cues, such as gazing and gestures, may also be helpful and aid listeners to more accurately predict the end of a unit. This explains why there is a greater degree of overlapping speech and a higher occurrence of speech repairs and disfluent speech in telephone conversation (Oviatt 1995).

Sophisticated methods of segmenting spontaneous speech at different levels have been developed by several research groups. Stolcke and Shriberg (1996) describe how word-level information can be used to segment speech into units. It has also been shown that prosodic cues can be used for detecting sentence boundaries (Stolcke et al. 1998). Hirschberg and Nakatani (1996) describe the relationship between discourse structure and intonational variation. Traum and Heeman (1997) examine boundary tones and pauses, and describe how they are related to grounding behavior in dialogue. Cettolo and Falavigna (1998) propose a method in which a combination of acoustic and lexical knowledge is used to detect semantic boundaries.

In studies of turn-taking in spontaneous conversation, another phenomenon has been observed. A speaker who utters a certain kind of phrase in conversation, such as a greeting, question or apology, typically expects the person he is addressing to come up with an appropriate response. Pairs of utterances such as question-answer or greeting-greeting are called *adjacency pairs* (Schegloff and Sacks 1973; Levinson 1983). For example, when person A encounters person B on the street, they exchange the following phrases:

A: How are you?

B: Not too bad

As suggested by the term adjacency pair, the second part of such a construction often follows immediately upon the first, as in the greeting-greeting example above. Nonetheless, the second part need not follow directly on the first, as in the example below:

A: May I have a bottle of Mich?

B: Are you twenty-one?

A: No

B: No

(example from Merritt 1976, quoted in Levinson 1983:304)

Adjacency pairs are important for the structure of spontaneous dialogue, since they are part of the communicative flow. When a person asks someone a question, he expects an answer to follow unless some problem has occurred or the listener is trying to make some sort of statement by not uttering the pending part of the adjacency pair.

3.5 Speech Acts

In an influential lecture, the philosopher J. L. Austin argued that utterances in a dialogue can be seen as *actions* performed by the speakers (1961). This can be clearly seen in a category of utterances which Austin call *performatives*, which spoken in the right context have the impact of committing the speaker to a certain course of action:

> "Since you won't allow me to work in research, I resign"
>
> "I bet you 500 crowns Linda won't finish her thesis this year"

Austin called these actions *speech acts*, and widened the concept by including not only performatives, but a much larger set of utterance types. According to Austin, any utterance has what is called an *illocutionary force*. For instance, when a speaker says: "Be careful – those steps are slippery!," the utterance has the illocutionary force of warning the listener.

John Searle (1965; 1979) later modified Austin's taxonomy, and suggested that one or more of the following speech acts can be assigned to all utterances:

- Assertives: Commit the speaker to something being the case
- Directives: Attempt by the speaker to get the listener to do something
- Commissives: Commit the speaker to some future course of action
- Expressives: Expresses the psychological state of the speaker
- Declarations: Bring about a difference in the world by an utterance

Analyzing human conversation with respect to speech acts can help us learn more about the general organization of spontaneous discourse. It then becomes clear that conversation consists of certain recurrent structures based on the utterances' illocutionary force. The specific speech act structure may vary depending on the task and domain of the dialogue, and its degree of formality. Speech acts are related to adjacency pairs in the sense that commonly occurring pairs like question-answer make up two subgroups in the much larger array of speech act categories.

Assigning speech acts to utterances in spontaneous conversation is not unproblematic. One and the same utterance may have different illocutionary force depending on the context in which it occurs. Furthermore, for some utterances, the surface form does not correspond to what they are used to express in conversation. For instance, an utterance such as “Could you close that window” has the surface form of a yes/no question, while it has the illocutionary force of a request. Politeness norms require us to use the grammatical construction “Could you…” in such contexts, when what we want to express is in fact “Close the window right now.” Other examples are statements such as “You are ready to leave now,” which given the right prosody should be interpreted as a question or a command. Such utterances, which have been called *indirect speech acts*, can be problematic for spoken dialogue systems and will be further discussed in Chapter 4.

3.6 Conversational Implicature

In conversation, an interpretation of an utterance relies on more than what is literally being said. Consider the following constructed exchange:

> **A**: What time is it?
> **B**: There goes Kant on his morning walk

If interpreted literally, B:s response looks like an answer to a different question than the one A asked. However, since both speakers know that Kant is a very punctual man and takes his walk at the exact same hour every day, the answer is in its context both cooperative and informative.

Most of the time, we assume that our interlocutors want to convey a message to us and try to infer what they are saying when the literal meaning of their utterance fails to explicitly do so. According to Grice’s (1975) theory of conversational implicature, a set of rational rules (maxims) govern our conversational interaction. The overarching rule is the *Cooperative principle*:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged (Grice 1975:158-9).

What enables listeners to infer more than what is literally being said is that conversation is guided by the cooperative principle and the four subprinciples, or maxims:

- Maxim of quantity: Be exactly as informative as required
- Maxim of quality: Try to make your contribution one that is true
- Maxim of relevance: Be relevant
- Maxim of manner: Be perspicuous

According to the maxim of quantity, a person's contribution to a conversation should be neither more nor less informative than required at a particular stage in the conversation. Normally, we expect people to convey the informative content of their utterance in an efficient and straightforward manner. Therefore, when a person is being overinformative we may suspect that he or she is trying to say something other or 'more' than what the utterance literally expresses (Grice 1975). Consider the following fabricated example:

A: Did you wash the dishes?

B: I washed the dishes, vacuumed the floor, took the dog for a walk, went shopping and called your mother

When a person obviously and overtly breaks one of the maxims for some linguistic effect, Grice calls this *flouting*. The maxim of quality states that our contributions in dialogue should be truthful. When someone flouts this maxim, we may infer that he is being ironic. Here is a translated example of flouting the maxim of quality in human–computer interaction, taken from the AdApt corpus:

System: I did not understand that

User: You are one smart system

The theory of conversational implicature provides an explicit account of how humans are able to interpret dialogue contributions in context, and how it is possible to say 'more' than words express.

To briefly sum up, human dialogue is characterized by a number of specific features:

- Spoken dialogue is the most basic and fundamental form of human communication, and there are refined strategies for handling human conversation
- Language is a collaborative process, by which speakers and hearers make a united effort to make themselves understood
- Turn-taking in spontaneous conversation is a structured process
- Language is used to perform actions, and utterances in dialogue can be interpreted in terms of what the speaker intends to execute
- In conversation, the interpretation of an utterance relies on more than its literal content

Chapter 4

Linguistic Adaptation

In this chapter, we move from the general to the specific by addressing the topic of how users adapt their language in different dialogue contexts. Here, the terminology and theoretical framework presented in Chapter 3 is used as a background for investigating how speakers adapt their spoken language to human and non-human interlocutors. The chapter is divided into sections, each representing a separate level of linguistic adaptation. For each section, strategies used in human–human dialogue are first reviewed. Subsequently, adaptation in human–computer dialogue and its implications are discussed. In the final part of the chapter, we attempt to summarize Chapters 3 and 4 by discussing how our understanding human–human communication can help improve future human–computer interfaces.

In human–human dialogue, linguistic adaptations occur as speakers modify their speech in accordance with several different factors:

Who is speaking – whether the speaker is an adult or child, if he is the boss or an employee, with an extrovert or introvert personality, as well as his experiences from previous dialogues

Who is listening – the status of the relationship between speaker and listener, and the speaker's beliefs about the listener's linguistic capabilities (e.g. child/non-native/hearing impaired/elderly)

What the dialogue is about – whether the topic is personal or professional, and whether the dialogue is social or task-oriented

Where the dialogue takes place – in a quiet or noisy environment, if it is a face-to-face dialogue or on the telephone, in a public or private setting

How the dialogue proceeds – if the dialogue runs smoothly, or if there are instances of misunderstanding or nonunderstanding among the dialogue participants

In human–computer interaction, an important question is how the users view the computer as a 'listener'. When people engage in interaction with a spoken dialogue system, their linguistic adaptations to some extent reflect their preconceived opinions about the linguistic capabilities of the computer. These opinions are based on a number of different factors:

- Previous experience with computers, and the particular task at hand
- How spoken human–computer interaction is described in popular culture, such as books and movies
- Their own experiences of using speech interfaces
- The general level of technical advancement in society at the time of their interaction
- The current capabilities of spoken dialogue systems

As mentioned in the previous chapter, spontaneous human dialogue is an adaptive process by which the speaker and listener attempt to make their interaction as efficient as possible. Spoken language has been around for a long time, and speakers use a variety of strategies for making conversation work as smoothly as possible. In conversational settings, language users continuously monitor their speech and the effect it has on listeners to ensure that they are making themselves understood. Depending on whom we are speaking to, and the context of the dialogue we are engaging in, we try to make sure our manner of speaking is in correspondence with the addressee's expectations and needs.

4.1 Phonetics and Prosody

Speakers modify the pronunciation, phrasing and prominence of their speech to communicate efficiently. For instance, we often speak quickly and use reduced language forms when talking to family members or close friends. On the other hand, we tend to speak more slowly and carefully when addressing non-native or elderly listeners. Phonetic and prosodic adaptation is a means of attracting the attention of young or elderly listeners or facilitating the comprehension for listeners with special needs. The status of the information conveyed (given or new), the present state of the grounding process, and the speaker's attitude towards the listener are other factors that contribute to how people adapt their manner of speaking.

4.1.1 Adaptation in Human Dialogue

At the phonetic and prosodic levels, the adaptation to the listener is manifested as modifications of the speaker's pronunciation. Lieberman (1963) observes that both the acoustic realization and the auditory perception of a word is a function of the semantic and grammatical context in which it occurs. Subjects were asked to read meaningful grammatical English sentences at a fast speaking rate. Some of the sentences consisted of maxims or idiomatic expressions, while others were less predictable in terms of content. The results of the experiment indicated that words which occur in predictable contexts can be seen as redundant, and are likely to be less carefully pronounced than words which occur in other contexts. For example, the articulation of the word *nine* in the idiomatic expression "A stitch in time saves *nine*" differs from how it is pronounced in the utterance "The number that you will hear is *nine*." In the latter context, the speaker will carefully articulate *nine* since this is the part of the utterance that carries the most information (Lieberman 1963).

When the acoustic signal is insufficient, semantic and pragmatic sources of information can help the listener fill in the blanks. Content words, which contain information units necessary to understand an utterance, are likely to be pronounced more carefully than function words. Van Bergem (1993) notes that for reasons of efficiency, speakers pronounce only the important parts of a message clearly and assume that their listeners are able to infer the rest. In a study examining acoustic vowel reduction, he found that sentence accent, word stress and word class significantly effect formant frequency and vowel duration. It can be noted that while vowel reduction is often realized as a movement from a vowel phoneme into a schwa in English, this is not generally the case for languages such as Swedish or Dutch (Nord 1986; van Bergem 1993).

Lindblom's (1990) H&H (hyper and hypo) theory is based on the observation that speech production is adaptive. According to this theory, the variability of the speech signal can be explained by the complementary relationship between signal-dependent and signal-independent properties in communication. As listeners, we are aided by cues outside the speech signal itself when we interpret the speaker's output. Consequently, a speaker who addresses a close friend about a familiar topic in a quiet room tends to speak quickly and use a reduced pronunciation. However, when the signal-independent factors are less beneficiary it becomes more difficult for the listener to predict the content of the message and discriminate between alternative lexical choices. In this context, the signal-dependent phonetic aspects of the speech become more important. A speaker who addresses someone unfamiliar in a noisy outside environment will typically adapt to this listener's needs by slowing down or by making an effort to speak clearly (Lindblom 1990). Moon and Lindblom (1989) show that clear speech is not merely normal speech which is produced louder. When speaking clearly, articulatory gestures are reorganized and acoustic patterns changed.

Studies have shown that when adults direct themselves towards infants and young children their speech takes on certain features. Cross-linguistic data has confirmed that speakers who want to attract the attention of newborn babies use a higher mean fundamental frequency (f_0), widen the range of their f_0, use short utterances, insert long pauses and whisper (Bredvad-Jensen 1995; Sundberg 1998). This speaking style has been called *infant-directed speech*. When compared to normal speech directed toward adults, it has been shown that infants pay more attention to and respond affectionately towards speakers who use infant-directed speech (Sundberg 1998).

Speakers who address listeners with hearing disabilities often make an effort to speak clearly. In order to investigate this, Picheny, Durlach and Braida (1985; 1986) asked subjects to read fifty nonsense sentences in a conversational and clear manner, respectively. The sentences were then analyzed with respect to a number of acoustic and prosodic features. When compared to the conversationally pronounced sentences, the clear speech was substantially longer in duration and contained more and longer pauses. Furthermore, vowels that were reduced or modified in conversational speech were found to be carefully pronounced in clear speech. Stop bursts and most word-final consonants were also realized in the clearly articulated sentences. In addition, the intensity of obstruent sounds was greater in clear than conversational speech (Picheny, Durlach and Braida 1986). Experiments also gave evidence that listeners with hearing impairments who were subjected to the two sets of sentences found it easier to understand the clear speech. The average difference in intelligibility between the conversational and clear speech was 17% (Picheny, Durlach and Braida 1985).

4.1.2 Adaptation in Human–Computer Interaction

In human–human dialogues, adapting at the phonetic and prosodic level clearly pays off. By modifying our speech according to the listener and context, we are able to efficiently make ourselves understood. If someone appears to have difficulty hearing or understanding what we say, we automatically speak louder and articulate more carefully. Since these strategies are such an integrated part of our language abilities, it is not surprising they are employed in human–computer interaction as well. To some degree, humans simply cannot help hyperarticulating when someone appears to have difficulty understanding. In addition, speakers who engage in dialogue with computers often modify their speech to meet what they believe to be the limitations of the system. For instance, speakers sometimes insert pauses between words because they think this will improve the accuracy of the speech recognizer.

Contrary to what speakers often believe, phonetic and prosodic adaptations which result in hyperarticulate speech are currently not very effective when your conversational partner is a computer system. Most state-of-the-art speech recognizers are trained on 'normal' speech, and are not very good at handling spoken input which is either reduced or exaggerated in pronunciation. Soltau and Waibel (2002) report that a state-of-the-art speech recognizer was degraded in performance by 30% when processing hyperarticulated speech instead of normal speech. Attempts to instruct users on how to speak by urging them to 'speak naturally' have not been successful (Shriberg, Wade and Price 1992). Instead, future spoken dialogue systems will have to be able to handle users' phonetic and prosodic adaptations. This issue is further discussed in Chapter 5 below.

Oviatt and colleagues have studied phonetic and prosodic adaptations in computer-directed speech (Oviatt, Bernard and Levow 1998a; Oviatt, MacEachern and Levow 1998c). In experimental sessions, volunteer subjects interacted with what they believed to be a fully functional multimodal system. Subjects could interact with the system using pen and voice, and the purpose of the study was to examine how speakers adapt their behavior while attempting to resolve errors. As will be discussed in detail in the next chapter, speech recognition errors often elicit linguistic adaptations in subsequent user utterances. Oviatt et al. (1998a) used a semi-automatic method of dialogue management for the experiment, thus enabling the simulated errors to occur randomly and without the human operator's interference. In their analyses of user behavior during error resolution the authors found that speakers:

- increase linguistic contrast, by either modifying the lexical content of their input or changing from one input modality to another
- increase hyperarticulation, by increasing the number and total time of pauses, increasing the final falling intonation and speech segments
- decrease variability, by decreasing amplitude swings and pitch range

(adapted from Oviatt et. al 1998a)

Oviatt and colleagues have developed what they call the CHAM model (Computer-elicited Hyperarticulate Adaptation Model) to account for changes in users' speech during error resolution. The long-term goal is to develop a model that can be used to predict how users adapt their language and speech during error resolution (Oviatt et al. 1998a; Oviatt et al. 1998b; Oviatt et al. 1998c). The question of how users' linguistic adaptations are manifested during human–computer error resolution is addressed in Chapter 5 below.

In summary, human beings continuously adapt their manner of speaking to meet what they perceive to be the requirements of the communicative situation and listener. In human–human dialogue, to adapt one's manner of speaking is an efficient strategy for improving communicative success. In particular, people make an effort to speak clearly when addressing a listener with special needs, such as a young child or a person with hearing difficulties. As Oviatt et al. (1998c) observe, human beings who interact with a computer may view this conversational partner as a listener who is similarly "at risk." If the user's initial expectations are low, she may modify her manner of speaking immediately, just as one would do when addressing someone in a noisy environment. A user with higher initial expectations may come to adapt her speech if or when the human–computer dialogue has become problematic.

4.2 Lexicon and Syntax

During spontaneous dialogue, speakers may be uncertain about how to refer to objects in a way that will make it easy for their conversational partners to identify them. Similarly, speakers use different types of grammatical constructions depending on the task and dialogue partner. In the course of a human–human dialogue, we try to come to an agreement with our interlocutors on which terms and constructions to use. In human–computer interaction, the adaptive process is a different one. Instead of negotiating, the users try to figure out how to express themselves in order to make the system understand them. This adaptation, in contrast to human–human conversation, is more or less unidirectional. However, knowledge about lexical and syntactic adaptation in human dialogue is nevertheless useful for improving spoken dialogue systems, since it appears as if similar mechanisms are at work in both types of interaction.

4.2.1 Adaptation in Human Dialogue

In a series of experiments, Garrod and Anderson (1987) investigate conceptual coordination. Pairs of subjects were asked to participate in a maze game, and in the process they established principles of interaction operating at a local level. Subjects who repeatedly referred to the same objects came to converge into using the same terms, a phenomenon the authors term *lexical entrainment*. The authors suggest a local output/input principle, which states that a speaker formulates his utterance according to how he interpreted his interlocutor's last utterance. Speakers who adhere to this principle are likely to establish a mutually acceptable way of referring with a minimum of joint effort (Garrod and Anderson 1987).

Brennan and Clark (1996) report that two different groups of people who refer to the same common objects in separate conversations will not use the same terms in more than 10% of the cases. However, it has been shown that although variability is high between conversations, it is relatively low within a conversation (Garrod and Anderson 1987; Brennan and Clark 1996). According to Brennan and Clark (1996) a speaker who suggests a way of referring to something is proposing a way of conceptualizing that object. If the interlocutor agrees with this proposition, a *conceptual pact* is founded. In Brennan and Clark's experiment, speakers first talked about cards that depicted different types of shoes. Once they had come to an agreement on how to refer to a particular shoe ('the dress shoe'), they kept using this term in subsequent references even though they were later discussing cards with different objects including only one shoe. When a conceptual pact has been established, the authors argue, speakers are sometimes overinformative in subsequent references instead of introducing a new term.

Levelt and Kelter (1982) showed a convergence effect for question-answer pairs in spontaneous spoken Dutch. Speakers were asked "What time do you close?" or "At what time do you close?" Results showed that subjects often answered correspondingly, responding with "Five o'clock" to the first question and "At five o'clock" to the second. Speakers thus repeated the syntactic structure/lexical items from the question in their answer.

Branigan, Pickering and Cleland (2000) argue that both listeners and speakers have something to gain from coordinating their spoken dialogue at the semantic and lexical levels. Listeners benefit because there is a lesser risk of misunderstanding, and speakers benefit since coordination reduces their cognitive load. The authors then hypothesize that these benefits should also hold with respect to other linguistic levels. In an experiment, one subject was given scripted descriptions which were varied in terms of syntactic structure. The other speaker's choices of syntactic structures were observed to vary correspondingly. Participants can thus be said to coordinate their syntactic structures in dialogue.

4.2.2 Adaptation in Human–Computer Interaction

People modify their choice of vocabulary and syntax to match that of their conversational partners. Just as in the case of adaptations of manner of speaking discussed above, this accommodation can be an effective strategy in human–human interaction. Communication becomes easier once there is an established common terminology, and interlocutors are accustomed to collaborating as they find a mutually acceptable way of referring in the dialogue.

As in human–human conversation, people who interact with computers using natural language are affected by their conversational partners' lexical and syntactical choices. Unlike phonetic and prosodic adaptation, which (if exaggerated) can degrade the performance of current dialogue systems, lexical and syntactic adaptation can be an effective strategy in human–computer interaction. Most spoken dialogue systems are designed to be able to handle as input the same vocabulary and syntactic constructions produced as output. Consequently, users' linguistic adaptations can be helpful as long as they mirror the language used by the system. However, as discussed in Chapter 5 below, not all kinds of lexical and syntactical adaptation can be adequately handled by current spoken dialogue systems. If a word is not in the speech recognizer's lexicon, the user may try a similar word in the same category. However, it may be the case that this whole class of words (such as the capitals of Europe) is missing from the lexicon, resulting in a communicative breakdown. When the human–computer interaction becomes problematic, some speakers react by using command-like language or by producing long and syntactically complex utterances.

In a study using typed input, Kennedy et al. (1988) asked groups of subjects to interact with what they believed to be a human or a conversational computer system, alternatively. Systematic stylistic variations were found in the language input of those subjects who believed they had been interacting with a computer system. In the users' input, lexical choice was limited and few instances of pronominal anaphora were found.

Guindon, Shuldberg and Conner (1987) describe a simulated study in which subjects could type questions to what they believed to be a computerized advisor. Their results indicated that users were very limited in their syntactic constructions. Most sentences produced were short and simple, containing few examples of subordination, coordination or relative clauses. Even though pronouns are short and easy to type, less than 3% of all sentences were found to contain any such items. Users also appeared to believe that the natural language interface could not handle fragments.

In an experiment where users submitted spoken as well as written input to a computer, Zoltan-Ford (1991) found that it was possible to get users of dialogue systems to adjust the length of their natural language input string to match that of the system. Interestingly enough, these results were not found to be affected by the mode of communication, so that the effects were observed in spoken as well as written user input. Zoltan-Ford also observed that it appeared to be easier to influence people to submit terse than verbose input. Other studies using text-based systems have shown that subjects who think they are interacting with a computer tend to simplify the syntactic structures of their language, using short and simple utterances (Dahlbäck 1991; Jönsson 1993).

Cheepen (1996) reports on a study using a speech interface to a word processing task, in which users were asked to submit a simple "yes" or "no" answer to a confirmation prompt. This message was also displayed graphically on the screen in front of them. However, instead of merely responding with one of these words, subjects would often repeat the item that was being confirmed. It thus appeared difficult to constrain their input to match that of the system's requirements.

Gustafson et al. (1997) present results from two experiments with simulated spoken dialogue systems. Users were found to reuse the language which occurred in the system's spoken output, both in terms of lexical choices and grammatical constructions. The authors note that a system that appears to prefer complete sentences encourages subjects to reuse large parts of the system's question in their answer. A system that appears to only be able to handle simple structures, on the other hand, elicits elliptical user utterances.

Brennan (1996) investigated people's willingness to adopt the vocabulary of the text-based and speech-based computer system they were interacting with, respectively. Users interacted with a simulated dialogue system that sometimes answered with another term than the one proposed by the human interlocutor. The system sometimes answered by an *embedded* correction, as in:

User: what college does Aida attend?

System: the school Aida attends is Williams

Alternatively, the system answered by an *exposed* correction, as in

User: what college does Aida attend?

System: by college, do you mean school?

User: yes

System: the school Aida attends is Williams

Brennan predicted that people would be more likely to adopt the system's vocabulary after an exposed correction, but that embedded corrections would also have some effect. Her findings suggest that the system's lexical choices are likely to affect users' choices, and that people are as willing to adopt the terms of a computer as they are to adopt the terms of a human conversational partner. The patterns for the text-based and speech-based systems were similar, although the effects for the speech-based system appeared more temporary.

According to Brennan, the *lexical convergence* effect in human–computer interaction should be seen as corresponding to (but not identical with) *lexical entrainment* in human–human conversation. One obvious difference is that while two people can discuss which terms to use, a computer interface cannot be assumed to have the ability to negotiate (Brennan 1996). Since computer programs generally are not constructed to discuss their own terminology, entrainment in Brennan and Clark's (1996) sense is not possible in state-of-the-art human–computer interaction. Nonetheless, there is clearly a unidirectional influence by which the terminology of a natural language system is likely to have effect on the user's choice of vocabulary.

As could be seen above, early studies in the field were often performed using text-based dialogue systems. Even though the user interfaces of these systems differ from the ones examined in the current thesis, the studies are worth considering for the role they played in the development of natural language interfaces. When looking at these experiments in sequence, the following pattern emerges: Text-based and less advanced spoken dialogue systems tend to elicit more of 'computer-directed' speech. As will be discussed in Chapter 6, more advanced speech-based systems with human-like features, such as animated talking agents, raise user expectations. At least initially, these speech-based systems appear to encourage more conversational linguistic behavior.

4.3 Pragmatics

In a broad sense, pragmatics can be understood as the study of language in context. Apart from what is actually being said in a dialogue, a multitude of factors contribute to speakers' linguistic adaptations in real-world situations. Turn-taking in spontaneous conversation is an adaptive process, and dialogue participants modify their discourse and feedback behavior to that of the other participants'. Sociolinguistic circumstances, such as the speaker's economic or social status when compared to the listener, also affect how the interaction proceeds. Whether the dialogue situation is formal or relaxed also contributes to how speakers behave. During a job interview, for example, the applicant is likely to be more well-mannered and polite than when afterwards having dinner and discussing the interview with a friend. These pragmatic aspects are highly culture-dependent, so that what is considered appropriate behavior in one society and era may be looked upon as incorrect in another.

When interacting with computers, people also adjust their discourse behavior, feedback strategies and level of politeness according to the dialogue situation and interlocutor. For a human interlocutor who believes he is interacting with a machine, it may seem redundant to be polite. If the user does not expect the computer to be able to handle politeness phrases and social remarks he may actively avoid such language. On the other hand, if a spoken dialogue system responds to social interaction and appears to be able to handle politeness (and impoliteness) phrases, users may be encouraged to continue using such linguistic strategies.

4.3.1 Adaptation in Human Dialogue

Turn-taking behavior is a dynamic process, and the organization of spontaneous discourse depends on the context and setting of the dialogue. Levinson (1983) observes that although the number of speakers in a conversation may vary from two to twenty (or more), the process of taking the turn still appears to remain orderly. As discussed in Section 3.3, studies of grounding in dialogue have shown that participants continually give each other feedback on the state of the interaction. These feedback cues (e.g. 'ok', 'mhm', 'alright' or gestural equivalents) are also part of the adaptive process, since they signal understanding and non-understanding in the dialogue, see Chapter 5. Moreover, the distribution and frequency of disfluencies and self-repairs in spontaneous conversation, depend on the dialogue setting and context. As mentioned in Chapter 2, Oviatt (1995) showed that speakers who interact face-to-face are less disfluent than speakers who talk on the telephone.

Studies of institutional dialogues, such as business meetings or court proceedings, have shown that speakers adjust their language to reflect the fact that they are engaging in a particular type of interaction (Drew and Sorjonen 1997). The social context of a dialogue affects people's way of talking at all levels, including their choice of lexical items, syntactic structures and prosody. For instance, a doctor who gives her patient advice on how to improve his diet might use imperative constructions like "Put less sugar in your cooking," or the hypothetical warning "If I were you…" (Drew and Sorjonen 1997). These types of constructions can perhaps not be used by the patient's friend, at least not with the same impact.

Brown and Levinson (1987) discuss how variation in politeness strategies manifests the social world, and its order of things. The difference in power and social status between two speakers is reflected in their politeness behavior, so that the speaker who is lower in rank is usually more polite than the one who is ranked higher. Blum-Kulka (1997) observes that the Gricean maxims are subject to contextual variation. In institutional settings, how the maxims are interpreted and adhered to is in the hands of the person in power. For instance, when a question is asked in a classroom it is the teacher who decides when the student's answer is informative enough (Blum-Kulka 1997).

The purpose of all human–human dialogues is not to conduct business, or transfer information. On the contrary, the goals of social dialogues are interpersonal, such as getting to know the other participants. Even in more formal or transactional dialogues, socializing functions as a way of getting the interaction started. Cheepen (1988) reports that small talk can be used at various points in the course of a dialogue, to continuously build rapport and trust.

Speech Accommodation Theory (SAT) focuses on how speech shifts during social interactions (Giles et al. 1987). According to this theory, speech accommodations can be *upward* (a modification toward the form valued by society) or *downward* (a modification towards stigmatized language forms). In general, it is argued that speakers are motivated to modify their speaking style to achieve certain goals, such as getting their interlocutors' social approval and maintaining a social identity.

4.3.2 Adaptation in Human–Computer Interaction

Early studies in human–computer interaction suggested that pragmatic adaptations, manifested in the usage of turn-taking cues, feedback, disfluencies, politeness language, irony or indirect speech acts, rarely if ever occur in human–computer interaction.

Some speech-based systems with graphical user interfaces have employed click-to-talk buttons for user input, something which simplifies turn-handling by restricting user initiative (see, for instance, Gustafson and Bell 2000). Early telephone-based systems were sometimes designed with cues such as "speak now" or a beep to signal that the system had handed over the turn to the user. However, Wooffitt and MacDermid (1995) found that subjects sometimes disobeyed instructions by speaking before the beep. In particular, speakers failed to modify their turn-taking behavior in cases when the system had previously misunderstood their utterance. After a system misunderstanding, speakers were eager to initiate the repair as soon as possible (Wooffitt and MacDermid 1995). As a response to users' unwillingness to adapt to system-directed turn-handling in telephone-based systems, in particular during error handling, methods for allowing user barge-in have been developed (Ström and Seneff 2000). In spoken dialogue systems with an open microphone and barge-in, turn-taking is less limited but can also be more complicated to handle.

Shriberg (1996) compared human–human and human–computer interaction with respect to turn-taking and speech disfluencies. The author found that disfluency rates were lower in human–computer dialogues, and that they were especially uncommon in a turn-initial position. It is suggested that initial disfluencies in human–human dialogue are used to coordinate exchanges and regulate turn-handling.

Reilly (1987) reports that users of a text-based interface appear to be unwilling to use utterances that indicate that they need assistance, or signal misunderstanding in the dialogue. The dialogues lacked feedback as well as explicit requests for help, and the author reports that:

> "People seem reluctant to say 'OK', 'Right',
> 'I understand', and such things to a machine."
>
> (Reilly 1987:72)

Dahlbäck (1991) observes that subjects who interact with a text-based simulated system refrain from using indirect speech acts, politeness markers, feedback, and backchannelling items. Analyses of the text-based human–computer dialogues showed that the human interlocutors did not use social greetings, and that they sometimes rephrased their previous input or suddenly changed the topic of the conversation in a manner that in human–human conversation would be considered rude. In the example below, the user abruptly and without warning changes the topic of the dialogue after the system has asked for a clarification of the previous input:

User: How many subjects are taken at the same time?

System: Between 3 and 5 subjects at the same time

User: Which subjects?

System: Specify your question better

User: Examination

(Example from Dahlbäck, 1991:129)

As a criticism against Dahlbäck's (1991) line of reasoning, it can be argued that if the user in the example above seems somewhat impolite, so does the system. The user's command-like language is merely her way of responding to a system which appears to be designed to disregard the principles of human discourse. A human dialogue participant who asks for a clarification would most likely add the word 'please' or something similar, instead of bluntly saying "specify your question better." Furthermore, the dialogue system's rather unsocial behavior is manifested in its response to indirect speech acts:

User: Is there any cheaper wine?
System: Yes

(Example from Dahlbäck, 1991:130)

When confronted with input that has the surface form of a yes-no question but is in fact an indirect speech act requesting information about which wine is cheaper, the simulated system appears rather uncooperative. If the users of this dialogue system are unwilling to engage in social behavior and refrain from using indirect speech acts, it comes as no surprise.

In contrast to Dahlbäck's (1991) view that computer-directed speech will be characterized by 'unsocial' features, Nass and colleagues (Nass and Steuer 1993; Reeves and Nass 1996; Nass and Moon 2000) argue that humans cannot help but being social when they use certain kinds of interactive media and computers. People know that computers are not people, yet their behavior signals otherwise. Nass and Moon (2000) review a number of experimental studies, all of which demonstrate how people apply social rules and expectations to human–computer interaction without being aware that they are. For instance, stereotypes about gender and ethnic background are applied to computer interfaces. The authors argue that these strategies are *not* caused by a belief that the computer is essentially human (*anthropomorphism*). Instead, people are to a large extent unaware of their own behavior and apply social rules during human–computer interaction without even thinking about it.

Bickmore and Cassell (2001; forthcoming) investigate how people interact with an embodied animated agent, acting as a real-estate agent. Subjects engaged in conversation with either one out of two versions of the system; one which engaged in social dialogue, and one which did not. Results indicated that the social dialogue had an impact on the user's trust in the agent, in particular for those users who were categorized as being 'extrovert' (Bickmore and Cassell 2001). The authors find that users who are not pressured for time to complete a task are quite willing to engage in social dialogue with an animated agent, and that many subjects prefer this interactional style (Bickmore and Cassell forthcoming).

It has been assumed that people who interact with spoken dialogue systems will not use irony or meta-language, i.e. comments about the dialogue itself or the capacities of the system. However, conversational implicature and flouting of maxims have been shown to occur in human–computer interaction. Nass and Moon (2000) have suggested that the Gricean maxims and conversational politeness norms, which are frequently used, are more likely to be applied in human–computer interaction than more infrequent rules of conduct. According to Grice's maxim of relevance, a dialogue contribution should be relevant in the current context (Grice 1975). In human–computer dialogues, this maxim is sometimes flouted by both dialogue participants. When a spoken dialogue system fails to interpret the user's input, it is often the case that the speech understanding module comes up with a faulty hypothesis, which can result in an erroneous and seemingly irrelevant system response. The system thus appears to be 'flouting' the maxim of relevance, as in the example in Table 4.1:

Table 4.1 The August system 'flouting' the maxim of relevance

User	**August**
Hur är det att vara en dator?	Hyfsat
What is it like being a computer?	*Not bad*
Är du smart?	Nej då, jag skulle ändå vakna nu
Are you smart?	*Oh no, I was waking up anyway*

One factor contributing to the occurrence of flouting of maxims in users' computer-directed speech is that if the computer seems to disregard the maxims, the user might think he can do it too. Consider the following example of a user flouting the maxim of relevance, taken from the AdApt database:

Table 4.2 A user flouting the maxim of relevance with the AdApt system

AdApt	**User**
Ett område på kartan är markerat. Där finns det just nu 8 lägenheter till salu.	Sortera bort lägenheter byggda före år 1900
An area on the map has been highlighted. There you can find 8 apartments for sale.	*Take out apartments built before the year 1900*
Kan du vänta lite längre?	Be husse och matte att köpa en större hårddisk
Could you wait a little longer?	*Ask master and mistress to buy a bigger hard disc*

To some extent, the ironic remark in Table 4.2 can be understood as directed to the person(s) responsible for designing and developing the spoken dialogue system, rather than the system itself. Nonetheless, a less advanced system would probably not elicit such user behavior.

4.4 Summary and Implications

> "The fundamental ability of human beings to adapt to their conversational partners makes the whole human/computer enterprise possible […] the way a conversational partner represents itself and the style in which it responds influence how a speaker designs utterances for that partner. Thus, the natural constraints imposed by the act of conversing should be exploited so that a user will be painlessly constrained from falling off the edge of the interface."
>
> (Brennan 1990:404)

The previous sections of Chapter 4 aimed to provide an overview of linguistic adaptations in human–human and human–computer interaction. Together with the theoretical background outlined in Chapter 3, the current chapter attempts to give some indication of how future spoken dialogue interfaces can be improved.

We saw that people who engage in dialogue adopt partly different strategies depending on whether their conversational partner is a computer or a fellow human being. To a certain extent, this can be explained by the fact that human–computer dialogues lack the transient quality which is typical of human–human dialogue. Another contributing factor is that the turn-taking capabilities of today's spoken dialogue systems are still rather restricted. With current system limitations, human–computer dialogues are slower and less reactive than human–human dialogues. In addition, users often find that it is tedious to recover from misunderstandings when they are interacting with a spoken dialogue system. This correlation between errors and linguistic adaptation is developed in Chapter 5 below. Furthermore, there are limitations as to which topics are addressed in a dialogue with a computer. The setting of the dialogue system, the task at hand, the user's initial expectations and prior experiences are also factors that come into play as people interact with computers.

Studies of human dialogue have suggested that there are recurrent patterns in how linguistic adaptations are manifested, at different linguistic levels. Phonetically, speakers adapt their speaking style according to the listener, setting and predictability of content to make as little of an effort as possible while remaining intelligible. 'At-risk' listeners are treated with special care, and adults use a special sort of language when addressing children. At the lexical level, speakers negotiate about which terms to use and then stick with them. Similarly, the syntactic patterns of one speaker's utterances can rub off on another speaker. The structure of discourse follows subtle turn-taking rules, which let speakers know when it is possible for them to enter into a conversation. As speakers and listeners, we are also aware that language can be used to say something 'other' than what is literally being expressed. When used in the proper context, a speech act can be uttered to perform an action. Finally, linguistic manifestations of politeness are also subject to an adaptive process: if someone is being polite to us, we are likely to be civil in return.

Early studies investigating text-based as well as speech-based systems indicated that computer-directed speech was often syntactically simple, and contained few pronouns, backchannels or politeness markers. However, these characteristic features of computer-directed language are not indicators of some inherent human behavioral strategy. Instead, such adaptations can be seen as the user's strategy for handling the limitations of every system they were recorded with. If we view language as collaboration between two partners who try to establish and expand on a mutual common ground, it becomes clear that users behave they way they do because of the system's (limited) linguistic capabilities. While a system that is consistent in its terminology will make it easy for speakers to know what to say, a system that is uncooperative may be more likely to elicit impolite user utterances. The precise manifestations of linguistic adaptation that are discussed in this thesis can also be seen as users' reactions to the limitations of the systems the human–computer dialogues were recorded with. However, although their appearance will differ depending on the user, context and capabilities of the current system, patterns of adaptation that are founded in the fundamental structures of human discourse behavior are likely to occur in every spoken dialogue system.

When interacting with spoken dialogue systems, the human adaptive process is in a sense twofold: the speaker adapts to the current dialogue situation by applying well-known strategies from human–human dialogue, and at the same time adapts to what he believes to be the limitations of the computer system. We agree with Brennan (1990) in saying that the human ability to adapt in conjunction with the existing regularities in human linguistic behavior is what should make it possible to improve the naturalness of future spoken dialogue systems.

Chapter 5

Errors and Adaptation

As discussed in Chapters 3 and 4 above, users often adapt their language in accordance with principles of human–human interaction and/or to meet what they believe to be the constraints of a spoken dialogue system. These chapters serve as a background for understanding the fundamental mechanisms of adaptation in spontaneous discourse. In the present chapter, we move on to examine some specific dialogue situations that are likely to elicit linguistic adaptations in human–computer interaction. More specifically, this chapter addresses the question of how speakers adapt their language when a problem or error has occurred in their dialogue.

In the following, several aspects of miscommunication in human–human and human–computer dialogue are discussed. The aim is to provide an overview of how speakers react to errors in spoken dialogue systems. In Section 5.1, we briefly discuss the occurrence of errors in human language production and perception. Furthermore, we address some previous research in the area of miscommunication in human dialogue. Section 5.2 deals with miscommunication in human–computer dialogue. In particular, the question of how linguistic adaptations can be used to identify errors in spoken dialogue systems is addressed. Finally, we attempt to summarize the chapter by discussing how knowledge of users' adaptive strategies can be useful for improving error handling in human–computer dialogue.

5.1 Miscommunication in Human–Human Interaction

One of the fundamental features of spontaneous dialogue is that participants sometimes have problems understanding one another, but are able to recover from errors and subsequently continue their interaction. As discussed in Chapter 3 above, human discourse is a fine-tuned and complex process. When miscommunication occurs in dialogues between people, it is often because the demand for efficiency in communication causes some small disruption in this process, in turn leading the listener to misinterpret the speaker.

5.1.1 Patterns of Linguistic Errors

As mentioned in Chapter 3.2, the syntactic structures of spontaneous spoken language are quite different from the structures and patterns of written language (Miller and Weinert 1998). For instance, spontaneous speech is often fragmented and contains repairs and disfluencies (Shriberg 1994; Oviatt 1995; Heeman and Allen 1999). It is important not to confuse these regularly existing patterns of spoken language with errors of competence or performance. As part of spoken communication, nonetheless, humans make mistakes in their production as well as in their perception.

Linguistic errors often exhibit systematic and rule-governed patterns. At the phonetic level, for example, there is a tendency for certain consonant changes to occur repeatedly in spontaneous spoken language. Well known examples are so called 'spoonerisms' (keep a tape/ **t**eep a **c**ape) and 'anticipations' (also share/ **alsh**o share) (van den Broecke and Goldstein 1980). Anticipations, which have been shown to occur frequently in English, are believed to reflect the fact that the speaker is already planning the next part of an utterance while physically producing the first.

At the semantic level, it has been suggested that speakers who accidentally exchange a whole word for another often produce a co-hyponym ('red' instead of 'blue'), a complementary term ('husband' instead of 'wife') or an antonym ('early' instead of 'late'). However, a word that stands in a synonymous relation to the intended word is seldom or never produced as a slip of the tongue (Hotopf 1980). It can be noted that linguistic errors by no means are restricted to spoken interaction. Regular patterns in the occurrence of errors have also been reported in writing, gesturing and sign language communication (Fromkin 1980).

5.1.2 Handling Problems in Dialogue

In human–human dialogue, it is not unusual for one person to misunderstand or fail to understand what another person is saying. According to Grice's (1975) maxim of quantity, described in Chapter 3, a person's contribution to a conversation should be neither more nor less informative than required at a particular stage in the conversation. We generally expect people to convey the informative content of their utterance in an effective and straightforward manner. When someone is being overinformative we may suspect that she is trying to say something 'more' than what her utterance literally expresses (Grice 1975). Because human–human conversation in general is characterized by efficiency and terseness, misunderstandings sometimes occur. A number of strategies for dealing with these unavoidable occurrences of miscommunication have been developed.

Clark (1994) notes that problems arise in many situations in every kind of human–human dialogue. While some of the problems may cause disfluencies, such as repetitions, filler words and hesitations, to appear in the participants' utterances, most of them are solved without much delay and evident effort. According to the theory of grounding, see Chapter 3, speaking is a joint activity and problems which occur in conversation are usually solved using collaborative strategies. The actions of one participant are not independent of those of the other, and a problem cannot be solved independently by one dialogue partner either. Clark proposes a model by which dialogue participants use three types of strategies when they try to handle difficulties:

- They try to *prevent* foreseeable but avoidable problems
- They *warn* partners about foreseeable but unavoidable problems
- They *repair* problems that have arisen

(Clark 1994:245)

In Clark's view, we should expect human–human dialogue participants to prefer preventatives to warnings, and warnings to repairs, all other things being equal. For instance, a speaker who foresees a delay that cannot be prevented, may warn his interlocutors by inserting a filled pause such as a 'uh' or 'um'. The reason for this is the relative high cost of repairing serious problems that have already arisen in a dialogue, compared to the relative low cost of a delay or an extra (perhaps unnecessary) dialogue turn.

Hirst et al (1994) observe that there is a limit in the amount of information participants can make explicit to one another in the course of a dialogue. They try to compensate for this by providing each other with cues signaling understanding, non-understanding and misunderstanding. Prior expectations are important, since what is consistent with what dialogue participants believe is going to happen often goes unnoticed. Accordingly, miscommunication can sometimes occur in human dialogue without any of the participants even being aware of it. On the other hand, an unforeseen turn in the dialogue makes participants observant and quick to notice if a misunderstanding has occurred (Hirst et al. 1994).

It has been argued that miscommunication should be seen as a specific feature of the more general process of collaborative behavior. In experimental sessions, Traum and Dillenbourg (1996) asked pairs of subjects to collaborate on a task which consisted in resolving errors. Based on their analyses of the users' behavior, the authors outline a predictive model of miscommunication and grounding in dialogue.

5.2 Error Identification in Human–Computer Interaction

Unlike human interlocutors, current dialogue systems are not very good at interpreting the subtle cues which signal that a misunderstanding is about to occur. Those discourse signals described by Clark (1994) as preferred, namely preventatives and implicit warnings, are often difficult for speech recognizers to handle. Instead, systems must be quick to respond when human–computer miscommunication is manifested through speakers' explicit warnings or repairs. Error identification is the process by which the system is able to discover an occurrence of communicative failure. In this context, users' linguistic adaptations can function as valuable cues, indicating a problem in the human–computer dialogue. Once the error has been identified, the system may choose to deal with this situation in different ways.

One motivation for quickly identifying errors in human–computer dialogue is that they have a tendency to cluster. Levow (1998) compared original input to verbatim repetitions that were corrections of the previous misrecognition error. For one particular spoken dialogue system, results showed that the probability of a speech recognition error after a correctly recognized turn was 16%, but immediately after a recognition error had occurred this probability increased to 44% (Levow 1998).

In the following section, we present two examples of problematic dialogues in the August corpus. In the remainder of this chapter, we will focus on those system errors in human–computer dialogue that elicit linguistic user adaptations. Problems that are the result of uncollaborative user behavior, as exemplified in the dialogue example in Table 5.2, will not be discussed. Furthermore, we will disregard system errors that do not elicit linguistic adaptations, either because the user never noticed the problem or because she did not modify her behavior.

5.2.1 Dialogue examples

Table 5.1 below is an excerpt from a longer dialogue between a young girl and the August spoken dialogue system. As discussed in Chapters 7 and 8, the August system was not designed to handle children's speech, and many of the younger speakers found it difficult to get the system to understand them. When her first utterance in Table 5.1 was misrecognized, this girl simply moved on to another question. After yet another unsuccessful turn she repeated her previous question verbatim, this time slightly hyperarticulated and louder. The system responded this time, although giving her a different answer from the one she wanted. The user tries to clarify her request by means of a meta comment which the system then fails to interpret.

Table 5.1 A problematic dialogue sequence between a child and the August system

User	August
Hur gammal är du? *How old are you?*	Det där förstod jag inte *I do not understand*
Berätta någonting om dig själv *Tell me something about yourself*	Det där förstod jag inte *I do not understand*
Berätta någonting om dig själv *Tell me something about yourself*	Strindberg var även intresserad av att göra guld av bly *Strindberg was also interested in making gold out of lead*
Nej men om dig **själv** inte om Strindberg *No but about **yourself** not about Strindberg*	Det där förstod jag inte *I do not understand*

In spoken dialogue systems, errors may occur on many levels. An error can be caused by a system misrecognition or misinterpretation, or by a failure in the dialogue handling capabilities or natural language generation module. In other cases, an error can be caused by a human interlocutor who provides the system with inappropriate input. Users sometimes do this with the deliberate purpose of testing the limits of the system. This is exemplified in Table 5.2 below, which is a short interchange between August and a male user:

Table 5.2 An exchange between August and a male user

User	August
Goddag *Hello*	Hej *Hi*
Vad är meningen med livet? *What is the meaning of life?*	Hej, vad trevligt att just **du** kunde komma *Hello, I am glad that **you** are here*

Many problems in human–computer interaction can be avoided if the system's limitations are made explicit to the user. The system can for instance inform the user about possible input options at every stage of the dialogue, and continuously give feedback on its interpretations to the user. However, there is a risk that the user perceives the resulting interaction as repetitious and boring. In human–human as well as human–computer conversation, there is a trade-off between being effective and risking to be misunderstood on the one hand, and making sure to be understood and being overinformative on the other. As mentioned in Chapter 2 above, the spoken dialogue systems that are the object of this thesis were deliberately designed to elicit complex and 'difficult' user behavior. The purpose was to explore what users anticipate the next generation of systems to be able to handle.

5.2.2 Feedback

One way of identifying errors in human–computer dialogue is to continuously give the user feedback about the system's dialogue state and ask for confirmation. According to this method, the absence of confirmation or the presence of an explicit warning or correction from a collaborative user should be seen as an indication that something has gone wrong. The system will then simply go back to the previous turn in the dialogue and make another attempt. As exemplified in this made-up example, the system can provide feedback explicitly in a separate turn:

System: Where in Stockholm would you like to live?
User: I would like to live in the Old Town
System: The Old Town. Is that correct?
User: Yes

A problem of this approach is that the resulting interaction can be perceived as unnatural and time-consuming, since many extra dialogue turns are required (Kamm and Helander 1997). As an alternative, the feedback from the system can be implicit in the sense that it is embedded in the subsequent turn:

System: Where in Stockholm would you like to live?
User: I would like to live in the Old Town
System: How many rooms would you like in the Old Town?
User: Three rooms

However, this method of providing feedback is not unproblematic either. If the system has in fact misunderstood the user's previous input, error handling can become rather complicated:

System: Where in Stockholm would you like to live?
User: I would like to live in the Old Town
System: How many rooms would you like downtown?
User: Downtown?
System: How many rooms would you like?

If speakers were to voluntarily give the system feedback about what they perceive as the state of the dialogue instead, this would make the process of asserting that the system is correct (or signalling that it is incorrect) smoother.

System: Where in Stockholm would you like to live?
User: I would like to live in the Old Town
System: How many rooms would you like?
User: Three
System: I found two three-room apartments now shown on the map
User: Good. What can you tell me about the one near the church?

In human–human conversation, people often give each other positive and negative feedback throughout the course of a dialogue. These acknowledgements (or, as the case may be, the lack of the same) function to prevent misunderstandings and can be used to notify a dialogue participant about the risk of an upcoming problem. In menu-based human–computer dialogues in task-oriented settings, spontaneous user acknowledgements have been unusual or nonexistent. This is not surprising, since the constraints on the dialogue are very strict and users are encouraged to use a command-like language. However, as spoken dialogue systems become less constrained and encourage more conversational behavior, it will appear more natural for users to voluntarily provide positive and negative feedback in the course of their interaction.

Brennan and Hulteen (1995) have suggested that a spoken dialogue system can provide its users with context-sensitive feedback. The authors present a feedback model for human–computer interaction based on the theory of grounding, see Section 3.3. This model includes positive evidence, such as back channels and explicit acceptances as well as negative evidence of understanding when some problem has been detected. Depending on the context, the feedback from the system will be more or less explicit. A response to an utterance with a high confidence score may be an implicit backchannel (such as 'uh huh'), while a response to an utterance with a low confidence score will clearly state how the system interpreted the user's previous input (Brennan and Hulteen 1995).

In a study of a telephone-based system for accessing e-mail that provided its users with the opportunity to give feedback, Ward and Heeman (2000) show that about half of the users gave the system acknowledgements at least once and a third of the subjects used them frequently during their dialogues. If users voluntarily give positive and negative feedback to the system they are interacting with, this could be used to identify errors. However, a difficulty with this strategy has to do with the fact that the feedback utterances rarely occur in a separate turn and may be difficult to extract and analyze adequately. Even if we assume that the dialogue system can interpret a series of fragments, another problem with the feedback approach is that there seem to be large individual variations in the use of acknowledgements (Bell and Gustafson 2000; Ward and Heeman 2000). As reported in Section 10.4 below, some users provided the AdApt system with feedback in virtually every turn, while others rarely or never responded in this way. An error identification method that is based on this approach may turn out to be successful for some users and fail for others.

5.2.3 Phonetic and Prosodic Cues for Error Identification

As discussed in Chapter 4, users often adapt their manner of speaking as they interact with a spoken dialogue system. Furthermore, phonetic and prosodic adaptations often become accentuated in a situation where the human–computer dialogue has become problematic. If spoken dialogue systems are designed to recognize typical human strategies for avoiding or repairing occurrences of miscommunication, their abilities to identify problems at an early stage can be improved, thus avoiding long and complicated error sequences. As mentioned above, it has been shown that misrecognized utterances differ from correctly recognized utterances at the phonetic level (Levow 1998). Hirschberg, Litman and Swerts (1999) report that certain prosodic cues, such as f_0 variation, within-utterance silences and speech rate, are good predictors of speech recognition errors.

Previous studies with simulated systems have shown that utterances that are part of a corrective sequence often contain certain characterizing features. These features include increases in duration and pitch range, hyperarticulation and lengthening of inserted pauses (Oviatt et al. 1996; Oviatt and VanGent 1996; Levow 1998; Oviatt et al. 1998a; Oviatt et al. 1998b; Oviatt et al. 1998c). If the user's speech contains an abundance of such features, this is a clear indication that something in the human–computer dialogue has gone wrong.

Users of dialogue systems often adapt their manner of speaking when they encounter a problem, for instance by modifying their pronunciation toward hyperarticulation. In most cases, a user who adapts her speech does so in an attempt to make the system understand her better. However, in contrast to the user's expectations, exaggerated phonetic and prosodic adaptations elevate error rates of current speech recognizers. In other words, adaptations which are manifested as hyperarticulate speech can turn out to be counterproductive and worsen the human–computer miscommunication. The question then becomes how to handle this problem. One possibility would be to instruct users on how to modify their manner of speaking when interacting with the dialogue system, so that their speech better corresponds to the training model of the recognizer. However, this would make the human–computer dialogue less natural and put the entire burden of solving the problem on the user. As previously mentioned in Chapter 4, it has been shown that it is difficult to instruct users on how to speak by for instance telling them to "speak naturally" (Shriberg et al. 1992). An alternative approach is to accept that users adapt their manner of speaking and work at developing systems that are better at handling hyperarticulate speech. Soltau and Waibel (2002) present promising results in this direction. The authors report that by training a speech recognizer separately on hyperarticulate speech, they were able to reduce error rates by 9%.

5.2.4 Lexical and Syntactic Cues for Error Identification

Speakers adapt their language at several levels when their human–computer dialogues become problematic. At the lexical and syntactic levels, certain features of the user's language signal the state of the discourse.

Krahmer and colleagues (Krahmer et al. 1999; Krahmer et al. 2001; Krahmer et al. 2002) base their studies of error detection in human–computer dialogue on the theory of grounding. The authors analyzed 120 dialogues with a Dutch train timetable information system. The study focuses on the acceptance phase of the grounding process, where the speaker's contribution may either be met by signals from a dialogue participant that are positive ('go on') or negative ('go back'). The authors list a number of positive and negative cues, as manifested in users' responses to the spoken dialogue system:

Positive ('go on')	**Negative ('go back')**
Short turns	Long turns
Unmarked word order	Marked word order
Confirm	Disconfirm
Answer	No answer
No corrections	Corrections
No repetitions	Repetitions
New info	No new info

(from Krahmer et al, 1999; 2001)

Linguistic modifications at the lexical and syntactic level can thus be useful for identifying errors in human–computer interaction. The presence of explicit lexical cues such as 'no' or 'stop' should be relatively easy to interpret and identify in spoken human–computer dialogue. Correspondingly, a long and complicated user turn or the absence of one or several user turns can be a clue suggesting that a problem has occurred. Once identified by a spoken dialogue system, an error can often be resolved. Just as in human–human dialogue, however, the adaptive behavior of the participant who had difficulties making himself understood sometimes persists a while after a problem has been handled.

5.2.5 Pragmatic Cues for Error Identification

Users' linguistic adaptations during error handling at the pragmatic level have so far been little explored. However, the occurrence of certain types of meta-language can be an indication that a problem has occurred in the dialogue. In a study which aimed at examining how prosodic cues could be used to identify speech recognition errors, Hirschberg et al. (1999) noted that certain words (such as 'help') indicated that an error had occurred. Similarly, Krahmer et al. (1999) showed that explicit disconfirmations (such as 'no') were much more common in the users' input after problematic dialogue turns than after non-problematic turns. In Section 10.4, we discuss how negative feedback in the AdApt system could be used to identify errors in the dialogue.

Social user behavior, or the lack of the same, can also be used to identify errors in human–computer dialogue. As discussed in Section 9.1, analyses of the August corpus suggested that users socialized with the animated agent, tried to flatter him and sometimes modified their language to reflect their disappointment with the system's answers (Gustafson and Bell 2000). For instance, speakers sometimes used insulting expressions, meta-language and ironical remarks during error handling. Users who explicitly insult a spoken dialogue system hardly do so because they believe the system will interpret or respond to their input. Instead, the use of swear words and expletive expressions can be seen as a way for the users to express their frustration or anger with the system's (lack of) capabilities.

5.2.6 Strategies for Error Recovery

Once an error has been correctly identified, we should assume that both the user and system would strive to recover from it. Error recovery is the process by which the system attempts to correct and repair problems that have occurred in the dialogue. Depending on where in the system the error has appeared, different approaches have been proposed. For speech recognition errors, it has been suggested that words that were originally misrecognized could be removed from the system's processing at the same time as the user is asked to repeat them, a strategy called "repetition-with-elimination" (Ainsworth and Pratt 1992). However, this approach requires confirmation from the system at every stage in the dialogue which easily becomes tedious. Instructive error messages from the system can also help the recovery from an error. As reported by Hunnicutt et al. (1992), users can be made to rephrase their input and thus recover from errors after having received an error message. Instructive error messages must be carefully designed to avoid increasing the difficulties or even causing new errors to appear. It is disturbing to have error messages appear repeatedly, and users may even prefer to let the problematic interaction go on for a few extra turns until the error is resolved.

In a study of errors caused by misrecognition, Smith and Hipp (1994) propose that verification subdialogues should be used selectively to recover from errors. The dialogue system should ask the user to verify his previous input, but only when the confidence score is low and the system has some indication that a problem has already occurred. In this way, a misunderstanding can be prevented as soon as it occurs instead of being allowed to continue for several turns. Smith (1998) has suggested that the context of the utterance is helpful in selecting which utterances to verify. It is important to be aware of the trade-off between exposing the user to unnecessary and tiresome verification subdialogues and the risk of getting into a very serious and perhaps unrecoverable problem in the dialogue (Smith and Gordon 1996).

The choice of error recovery strategy must depend on the sort of application or task the system is designed for. Within one and the same domain and a single system, the occurrence of an error can be more or less serious depending on where in the dialogue it appears. Kamm and Helander (1997) report on attempts in which the system is designed to consider the cost and probability of errors. Where the cost of system errors is very high, the user should be encouraged to switch to an alternative modality or even be redirected to a human agent. Several recent studies have described approaches in which the user is encouraged to switch to an alternative input modality as a means of recovering from errors (Oviatt et al. 1996; Oviatt and VanGent 1996; Suhm and Waibel 1997).

Finally, when no other option is available the system should offer the user a gentle way of ending the interaction. A user's final impression of a spoken dialogue system that has given him a hard time may perhaps be a little improved if at least the exit from the system is a civil one.

5.3 Summary and Implications

This section summarizes the chapter on errors and linguistic adaptation in dialogue. As in the final part of the previous chapter, some implications for the improvement of future spoken dialogue systems are also discussed.

In this chapter, we have seen that errors occur in both human–human and human–computer dialogue. In human dialogue, there are well-established strategies for handling upcoming problems in conversation. Human interlocutors give each other feedback, making it possible to prevent serious problems in the conversation. However, the need for terseness and efficiency in human communication sometimes causes unavoidable misunderstandings to occur. When resolving misunderstandings in dialogue, the well-known structures of human discourse are there to fall back on. Human interlocutors are able to draw upon their knowledge of how other people typically react to a dialogue misunderstanding. Linguistic adaptation strategies, for instance at the phonetic level, are often effective in human–human conversation. Thus, a speaker can be quite certain of how a human dialogue partner will respond to a particular type of adaptation and employ the means necessary for resolving the problem.

In human–computer interaction, misunderstandings and errors have a much greater impact and the potential to be more harmful. First of all, it is difficult for a computer system to respond to subtle human discourse signals which warn of an upcoming problem. Errors can thus seldom be avoided. Once an error has occurred, the system cannot always detect it at once. It can also be hard for a human interlocutor to understand why the system did not respond as expected to a specific utterance. During error handling, speakers often adapt their language at several levels. In contrast to the effect of such adaptations during error handling in human–human dialogue, user's linguistic adaptations sometimes worsen the problem in human–computer interaction. Hyperarticulate speech and inserted pauses between words degrade the performance of most speech recognizers, and long and syntactically complex utterances can be difficult for the system to handle. However, linguistic adaptations function as important cues for identifying problems in human–computer interaction. In conjunction, users' phonetic, prosodic, lexical, syntactic and pragmatic adaptations are useful indicators of miscommunication in human–computer interaction. Hopefully, future systems will be able to better handle these features, resulting in fewer problems, improvements in error resolution abilities and increased user satisfaction.

Chapter 6

Empirical Methods

This chapter deals with empirical methods, and some of the problems which need to be addressed when using these methods in spoken dialogue system research. In the first part of the chapter, we present arguments for using an empirical approach in human–computer studies in general and in the design of spoken language interfaces in particular. Next, we consider some relevant issues in the area of interface design and user expectations. In the second section of the chapter, the two types of systems and methods primarily used for data collection with spoken dialogue systems are discussed and compared. Subsequently, some of the specific problems of collecting, transcribing and analyzing spontaneous speech data are discussed. We then consider some ethical questions which are necessary to address when collecting and analyzing speech data. Finally, the chapter is concluded with a brief summary.

6.1 Why Empirical Methods are Necessary

Human–computer interaction (HCI) is a research field which describes all aspects of the interplay between a human being and computer (Helander, Landauer and Prabhu 1997). HCI studies include work in the areas of computer science, psychology, linguistics, ergonomics, sociology, arts and design. This thesis, however, deals only with a specific type of human–computer interaction: spontaneous face-to-face conversations with animated agents in spoken dialogue systems.

It has been shown that the best way to develop a human–computer interface is to empirically test and evaluate human behavior in different settings (Day and Boyce 1993). All the same, business developers and others working outside the field of HCI sometimes view user studies as time-consuming and expensive. Instead, the argument goes, the design of a natural language interface requires no more than a few moments of contemplation and intuitive reasoning. Since we are all used to speaking to other people and daily participate in dialogue, any person capable of a little introspection is equally well equipped for the job of developing a spoken language interface.

The above argument, known as *the egocentric intuition fallacy*, has been criticized by researchers in the field of HCI (Landauer 1997). It is based on the misconception that we as human beings have a clear and direct access to how our own minds function and operate. However, it is impossible to intuitively 'know' how to structure an interface in order to make it optimally usable and accessible. Moreover, as human beings we are often inclined to believe that what is true for ourselves must also be true for others. Because of limitations of imagination and experience, we believe that other people argue like us, reason like us, and speak like us. Nonetheless, human behavior is more variable than most of us can even imagine. Studies have shown that the individual differences in human–computer interaction are large. For some tasks, the variability in performance of equally trained subjects in terms of completion time is as high as 10 to 1 (Egan 1988).

Researchers in the field of HCI sometimes encounter resistance when they suggest empirically testing different ways of designing a user interface. In the development of commercial applications where there is pressure to be cost efficient, this may be because there is a lack of understanding about why the interface specialist cannot simply offhand produce the specification for how to build the system. Nonetheless, implementing a dialogue interface based on the designer's intuitions can turn out to be very expensive, since it may result in dissatisfied customers and the need to redesign the system. In the long run, systematic empirical testing and iterative development of human–computer interfaces are both cost-effective and efficient.

6.1.1 Designing Speech Interfaces

As pointed out by Zoltan-Ford (1991), one of the problems of natural language interfaces is the freedom they allow their users. How to handle the large individual variability in written and spoken input remains one of the big research challenges. In an article describing this as *the vocabulary problem*, Furnas and colleagues note that the probability for two users to spontaneously choose the same term in a text-based system in an application-related domain is less than 20% (Furnas et al. 1987). As a solution, it is suggested that a high number of alternative terms are made available to make this interface functional from the user's point of view. In a text-based system, the active use of a large vocabulary is relatively unproblematic.

For speech-based systems, the vocabulary problem is more complex. The larger the lexicon of the speech recognizer, the greater the risk of errors and subsequent problems in the dialogue. Yankelovich (1996) observes that desktop graphical user interfaces were developed to make the computer's functionality visible for the user. In contrast, speech-based systems hide their functionality from the users, who have no way of assessing the boundaries of what they can or cannot say (Yankelovich 1996). In the first generation of spoken dialogue systems, the speech recognizers' inability to handle out-of-domain words made it necessary to restrict the number of active words and use a system-driven dialogue structure. As a way of handling these limitations by restricting the user's options at every step of the dialogue, a menu-based approach was sometimes used. Today, menu-based interfaces are mainly used when errors must be avoided at any cost, as in banking applications. According to this method, the system explicitly tells the user which vocabulary items are available at each step of the dialogue:

System: Please say collect, calling card, third number, person to person or operator, now.

User: Calling card

System: Please enter or say your calling card number, now.

User: 908 949 1111 0000

(Boyce 1999:38)

Menu-based systems can be useful for restricted tasks, but tend to be frustrating because of their inflexibility. Moreover, these interfaces often require the user to employ a special kind of language appropriate for each particular system. As Boyce (1999) notes, menu-based systems are also rather inefficient since the user cannot state all of the information in a single turn, as in "I'd like to place a calling card call using the number 908 949 1111 0000." As mentioned in Chapter 2, the possibility to avoid going through a complicated menu-system and immediately express a specific request is often seen as one of the main advantages of spoken dialogue interfaces from the user's point of view.

Speech interface designers have long argued that it is necessary to allow speakers to interact using their own spontaneous language. Recent developments in speech recognition and understanding have made it possible for speakers to say almost anything they want, as long as their interaction is restricted to a narrow domain. In the last five years, systems for call routing and simple single-domain dialogue systems have been developed and publicly launched. Large data collections in the *How May I Help You* – system, developed at AT&T (Gorin, Riccardi and Wright 1997; Langkilde et al. 1999), have enabled 90% of the caller's initial requests to be accurately routed to a specialized operator or automatically handled. In *JUPITER*, a telephone-based speech interface to a weather forecast service developed at MIT, about 100,000 calls have been received since 1997. This system achieves word accuracy for novice users and in-domain utterances in 89% of all cases, and correct understanding in about 80% of all cases (Zue 2000).

6.1.2 Interface Design and User Expectations

The design of the user interface is one of the factors which contribute to the speaker's expectations of a speech-based system's capabilities. Other factors, such as prior experiences with spoken dialogue systems, educational background, general attitudes towards new technology etc., also affect the user's initial expectations. In the course of the human–computer interaction, these expectations are either fulfilled or not, depending on the success of the dialogue.

For telephone-based systems, it has been shown that the greeting prompt has a major impact on the user's expectations and general impression of the system (Boyce 1999; Edgington, Attwater and Durston 1999). To test user reactions to different system prompts, Boyce (1999) set up an experiment in which four styles of output were used. The prompts were either casual or formal, and the system either referred to itself using first person personal pronouns ("I" and "me") or not. For example, a casual prompt with a first person pronoun would be "Anything else I can help you with?" while the corresponding formal prompt without the pronoun would be "If you have an additional request, please say it now." Results showed that the subjects were significantly more satisfied with the versions of the system that referred to itself using first person personal pronouns. Interestingly enough, 80% of the subjects had not even noticed that the system used "I." For the prompts with first person pronouns, subjects seemed not to care whether they were casual or formal, even though the formal prompts were longer.

Whether or not a computer system should appear human-like is a question in dispute. According to Shneiderman (2000), a spoken dialogue system should not try to attempt to "model or recognize complex human behavior." The author argues against endeavoring to create computer systems that simulate human social or emotional behavior. Boyce (1999) reports that older research in the field of human factors indicated that human-like systems would create unrealistic expectations, followed by negative user reactions.

When engaging in face-to-face dialogue, humans employ a number of non-verbal signals along with spoken language. In the last decade, we have seen the emergence of spoken dialogue systems that are *embodied*, either by an animated talking face or a full-bodied character (Cassell et al. 1999). It has been shown that users who communicate with an embodied interface enjoy their interaction more and spend more time with the system (Lester et al. 1997). However, there is also concern that embodying the interface will give users unrealistic expectations and anthropomorphize the system. On the other hand, it has been argued that interface designers can take advantage of the fact that users view these systems as possessing certain human characteristics. For instance, Bickmore and Cassell (forthcoming) report that subjects who interacted face-to-face with an embodied agent had significantly less speech disfluencies than subjects who interacted with the same system on the telephone.

To sum up, the design of the user interface is one of the factors which contribute to shaping the user's expectations during human–computer interaction in a spoken dialogue system. While it has been argued that human-like interfaces can create unrealistic user expectations, it also appears as if users prefer a system that has an embodied human-like persona and refers to itself using a first person "I."

6.1.3 The Iterative Design Process

According to Day and Boyce (1993), usable human–computer interfaces will only evolve when developers employ a user-centered design process. The authors summarize the human factors principles for designing usable human–computer interfaces. These principles include an early and continuous focus on users and as well as an *iterative* design. This means that the user interface design process should be based on continual empirical testing. The process of testing and modifying the interface is repeated until the system "clearly meets users' needs for functionality and usability." The authors also emphasize that the tests can be conducted using prototypes when real systems are not available, but that an actual system should be used once it is possible (Day and Boyce 1993).

Landauer, who is also an advocate of iterative design, writes that "the best strategy for good design is to try various options […] test them, and be guided by the failures, successes and comments garnered in watching their use, redesign trying new options, and iterate." (1997:214). As the author points out, it may not always be necessary to go through a great number of iterations.

The best practice for the development of a spoken dialogue system thus seems to be the following:

1. Develop a simple functional system or simulated version of the system
2. Collect user data with the system
3. Analyze the collected data of user interaction
4. Develop a new version of the system, based on the collected data
5. Collect data with the improved system
6. Analyze the second set of data
7. …go back to 2 and repeat the process

Because of limitations of time and resources, it may be difficult to go through the necessary number of iterations in the development of a single spoken dialogue system. However, as will be shown in Part II below, lessons learned in the development of one spoken dialogue system can also be useful in the design of the next. Some aspects of system design and user behavior in human–computer interaction are more general, and should be useful in the innovation of the next generation of systems.

6.2 Methods of Data Collection

Until the last decade, controlled user experiments in laboratory environments were the only settings in which it was possible to collect data with spoken dialogue systems. As soon as speech recognition systems were robust enough, alternative and more cost-efficient methods were explored. As mentioned above, telephone-based information systems have been used in several countries around the world to collect large quantities of speech data, which has subsequently been analyzed and used to improve the performance of future systems. However, conversational spoken dialogue systems with animated agents capable of handling a complex domain (or more than one domain) have yet to be subjected to large-scale field trials.

In this section, we contrast two methods of collecting data and two frameworks for developing systems for data collection.

6.2.1 Wizard-of-Oz or Functional Systems?

This section compares two different types of systems which can be used to collect spoken dialogue data. Arguments for simulating speech interfaces are first presented, after which advantages of fully functional systems are listed. Sections 6.2.1 and 6.2.2 are interrelated in the sense that laboratory experiments can be used for either setting up simulated systems or for testing fully functional spoken dialogue systems. Stand-alone systems in public environments, on the other hand, can only be used at a large scale to test functional spoken dialogue systems. Fully functional systems can be upgraded and improved iteratively, as discussed above, but their users cannot yet improve them on-the-fly.

Fraser and Gilbert (1991) list some of the motivations for performing studies in which a human operator simulates all or parts of the system's capabilities, so called *Wizard-of-Oz* studies (Gould, Conti and Hovanyecz 1983). Despite recent improvements in speech recognition and understanding, as well as developments in the areas of dialogue management and natural language generation, the basic problem remains: designers of spoken dialogue systems cannot base these systems' natural language understanding capabilities on intuition or previous experience alone. Relying on human–human dialogue data is better, but still far from sufficient. For one thing, it is difficult to know which aspects of the human–human dialogue context and situation that are relevant in the development of a human–computer interface. Furthermore, because of their prior expectations regarding the capabilities of spoken dialogue systems, speakers may use partly different strategies with computers than those well-known from human communication. Since natural language interfaces are so complex, the general problem which Thomas (1995) calls "testing for technologies that do not exist" is particularly tricky in the development of spoken dialogue systems:

> "The designer is caught in a vicious circle – it is necessary to know the characteristics of dialogues between people and automata in order to be able to build the system, but it is impossible to know what such dialogues would be until such a system has been built."
>
> (Fraser and Gilbert, 1991:81)

When collecting data with a simulated system, it is important that the behavior of the system's wizard does not become too advanced and human-like. If it does, the resulting dialogues will perhaps not be very useful for the development of the functional system. As pointed out by Dybkjaer, Bernsen and Dybkjaer (1993), Wizard-of-Oz studies involve a trade-off between naturalness and keeping within the technological constraints of the system under development. In order to restrain the wizard's behavior, and to make the system's response time shorter, it has been suggested that a semi-automated method, e.g. with a predetermined template, is used (Oviatt et al. 1998a; Bell et al. 2000).

Subjects who interact with Wizard-of-Oz systems are told they are speaking with a computer. As Fraser and Gilbert (1991) observe, one of the necessary requirements of a simulated system is that the people who interact with the system really believe they are talking to a functional system. However, the researcher who has managed to convince his subjects that they are interacting with a computer is faced with a problem: Is this really justifiable from an ethical point of view? In accordance with the principle of information, listed in Section 6.4, the researcher is obliged to inform his subjects about the purposes of the study they are contributing to. With the Wizard-of-Oz method, the whole success of the experiment relies on subjects *not* being fully informed about the true purpose of their participation. Afterwards, when the dialogues have already been recorded, the researcher exposes the accurate version of the system setup. At this point, some subjects may feel they have been deceived and finish their session with a sense of disappointment.

Apart from the ethical considerations, there are other arguments for collecting data and analyzing user behavior using fully functional systems. Allen et al. (1996) have argued that unless we have real working systems, it becomes almost impossible to make fair evaluations of models and theories on human–computer dialogue management. Allwood and Haglund (1992) have pointed out that when engaging in Wizard-of-Oz dialogues, both parties are in a sense taking on roles in which they are not really representing themselves. For example, the wizard may play the role of travel agent, and the subject plays the role of the customer who is interested in buying a ticket. When people really work as travel agents, or really want help with their travel arrangement, their behavior might be quite different.

Early trials of the fully functional Philips automatic train time table information system (Aust et al. 1995) and the LIMSI RailTel system (Lamel et al. 1997) showed that making speech interfaces to real services publicly available was useful, since new groups of users were given access to the systems. Large amounts of data from users in real-life situations have been collected in several single-domain systems (Gorin et al. 1997; Lamel et al. 1997; Eskenazi et al. 1999; Zue 2000). While these systems have successfully been used to collect dialogues in limited domains, it could be useful to complement them by using fully functional systems that gather data from a wider range of users in several domains.

6.2.2 Controlled or Unsupervised Data Collection?

There are several advantages of performing controlled user experiments in a laboratory setting. First, when using this research method it is possible to begin by formulating a hypothesis about user behavior which can be systematically tested. If the results do not support the theory, the hypothesis has been falsified. Another hypothesis can then be formulated, a new experiment be set up, etc. Second, when undertaking a controlled study in a laboratory environment the researcher is able to gain a lot of background knowledge. The information concerning each subject's education, previous experience using speech interfaces, initial expectations etc., can be used to investigate any correlations that might exist between the user's behavior and the background data. Third, an experiment in a supervised setting has the advantage of minimizing the uncontrolled variables which influence the data in ways which are difficult to assess. Sources of disturbance in the acoustic environment as well as disobedient subjects can be eliminated so as to avoid contaminating the data. Consequently, it is easier to evaluate results from a controlled user experiment. A final advantage of setting up supervised experiments is that one can make sure that ethical considerations are dealt with in an adequate manner. When subjects have been called in to the laboratory, the researcher in charge of the experiment can inform them of the ethical guidelines. Before beginning their interaction, it is important to make sure all subjects are aware of the conditions of their participation, listed in Section 6.4 below.

Data collection in a publicly available stand-alone system which encourages spontaneous human–computer dialogue has its advantages, too. First, the dialogues collected with an unsupervised system are likely to be more natural and less constrained than the dialogues recorded in a laboratory setting. Speakers are less afraid to interact spontaneously, and do not care (or even know) about the expectations of the researcher behind the experiment. In this sense, corpora collected with a publicly exhibited system are likely to better reflect a 'true' picture of user expectations. Second, data collected in a publicly exhibited system comes from a wider range of users and thus has the potential of better representing the socio-cultural background of the speakers of this community. Depending on where and in what sort of setting the system is displayed, the demand for representativity is more or less difficult to meet. However, subjects who are recruited as participants in laboratory experiments do not always reflect the diversity of society and their view on the systems examined are perhaps not always representative. As Landauer (1997) notes, people who volunteer for studies on new technology are often people who like new technology. Third, since the researcher in charge of the experiment is not around when collecting data using an unsupervised system there is no risk of influencing the user inadvertently. In supervised experiments, there is always a risk that the user feels he or she is subjected to some sort of test, and has to do what is 'correct.' The downside of having a stand-alone system is that it is difficult to ensure that speakers are aware of the ethical terms regarding their participation, see Section 6.4. Signs informing users that they are being recorded can be attached to an unsupervised system, but there is no guarantee that all speakers have read and understood such a sign before starting their interaction with the system. This situation is especially difficult with young speakers who may not even know how to read. On the other hand, in a set-up such as the one used in the August system, see Section 7.1, it is virtually impossible to identify an individual speaker. This makes some of these ethical considerations less critical. A final advantage of collecting data in public is that because of the great expense of performing controlled user experiments, stand-alone systems can potentially collect great quantities of data at a low cost. Once collected, however, the data from a stand-alone system can be rather messy and difficult to transcribe and analyze.

A feature of human face-to-face interaction that distinguishes it from many other types of language use is that it is temporary and transient. Ordinary conversations are not usually recorded, and people say things in spontaneous conversation they would perhaps hesitate to write down. Clark (1996) notes that when people are aware of the fact that their speech *is* being recorded, as in the case of telephone answering machines, this immediately affects their manner of speaking. In order to meet the ethical requirements listed in 6.4 below, subjects have to be informed that the human–computer dialogue they are about to engage in is not an ordinary, transient conversation. When they are aware they are being recorded, speakers are likely to be more careful about what they say and how they say it.

When pros and cons have been weighed against each other, it appears as if advantages and disadvantages with both approaches can be found. Users behave more naturally and spontaneously when interacting with an unsupervised system in a public environment. The researchers' expectations are less likely to influence the users' behavior in a stand-alone system, and it is potentially possible to use such systems to collect large quantities of data at a low cost. However, it is difficult to systematically hypothesize and control for different variables when data is collected with an unsupervised system. Furthermore, it is sometimes difficult to make fully functional systems robust and advanced enough to collect interesting data. In addition, it can be problematic to get users' consent for collecting speech data in unsupervised systems. In particular, this issue is sensitive when children are recorded in public environments. Depending on the context and specific research question addressed, one or the other of the methods may be preferred. To get a full and complete picture of user's linguistic adaptations in spoken dialogue systems, it is probably useful to employ a combination of both methods.

6.3 The Annotation of Speech Corpora

In this section, we briefly describe the annotation of speech corpora, discuss some of the difficulties commonly encountered when analyzing spoken dialogue data, and some relevant aspects of the transcription and annotation of spoken dialogue corpora. In Part II below, specific details of the annotations of each of the corpora analyzed in this thesis are discussed.

6.3.1 Collecting and Labelling Spoken Dialogue Corpora

Written language corpora collected for linguistic analysis are often labeled at the grammatical level, with word class tags for every word token in the corpus. In addition, text corpora are for example labeled for syntactic functions, resulting in parsed treebanks. Other annotation schemes, which for instance include semantic information, are also used for specific purposes. Some of this labelling can currently be done automatically, using corpus tools and machine learning algorithms.

The collection and annotation of speech corpora is inherently more complicated, and requires several supplementary levels of linguistic analysis. Leech (1991) describes how in 1975, Jan Svartvik and his colleagues at Lund University made the corpus from the spoken part of the Survey of English Usage available in machine-readable form. The resulting London-Lund corpus of spontaneous spoken English (Svartvik and Quirk 1980) had a great impact on subsequent studies of spoken dialogue. In addition to the levels of analysis described above, the London-Lund corpus was hand-labeled at the phonetic and phonemic levels. The corpus was also annotated at the prosodic level, and analyzed at higher linguistic levels with tags for discourse markers and turn-taking behavior (Sacks et al. 1974; Leech 1991).

In the last decade, several large corpora of spontaneous speech have been collected and made available. For instance, the Edinburgh Map Task Corpus of task-oriented dialogues (Anderson et al. 1991), air travel information dialogues from the MADCOW group (Hirschman et al. 1992) and the Switchboard corpus of spontaneous telephone conversations (Godfrey, Holliman and McDaniel 1992) have made it possible to develop statistical language models and improved methods for analyzing spoken language.

6.3.2 Levels of Annotation

An annotated speech corpus can provide statistics for probabilistic language processing, something which is important in the development of spoken dialogue systems. Depending on what the data is to be used for, a corpus of spoken human–computer interaction can be analyzed at several or all of the following levels:

- Orthographic transcription
- Phonetic transcription
- Disfluencies and self-repairs
- Prosody
- Word class (parts-of-speech)
- Syntactic structure
- Speech acts and dialogue structure

Recent advances in the development of corpus tools have simplified the process of transcribing and analyzing speech data. Programs such as *Snack* (Sjölander 2001; Sjölander 2003) include an alignment algorithm that allows a speech file to be automatically labeled at the phonemic level given a verbatim transcription of the speech. At the prosodic level, ToBI (Tones and Break Indices) has been proposed as a standard for labeling English. This scheme captures features such as pitch accent, phrase accent and boundary tone (Silverman et al. 1992).

Even after a corpus has been labeled phonetically and prosodically, the occurrence of disfluencies and self-repairs can cause problems. Nakatani and Hirschberg (1994) report that self-repairs occur in about 10% of all spontaneous utterances. However, the authors show that acoustic and prosodic cues can successfully be used to detect speech repairs. Hindle (1983) describes a method of detecting speech repairs, in which the speaker goes back and repeats or rephrases parts of his previous utterance. Heeman and Allen (1999) observe that such repairs can make it difficult for a spoken dialogue system to interpret what the speaker intends to express, since they are no longer part of the speaker's intended utterance.

To be able to improve a spoken dialogue system's discourse capabilities, it is necessary to have speech data which is labeled at the speech act level. The goal is to be able to assign the illocutionary force intended with each user utterance. Jurafsky and Martin (2000) note that for most user utterances, this speech act interpretation is relatively simple. Statements commonly have the syntactic form of declarative sentences ("That is a nice apartment"), while imperative syntax is used for uttering commands or certain questions ("Tell me about restaurants in Stockholm"). However, as the authors note, there is no one-to-one mapping from the surface form to the illocutionary force of an utterance. In human–computer interaction, indirect speech acts can be complicated for the system to interpret. Nonetheless, there are regularities in how these utterances occur as well, something which annotation schemes try to capture. The DAMSL scheme is a system for annotating utterances in dialogue for illocutionary force and relationships between utterances (Core and Allen 1997; Stent 2001). Stolcke et al. (2000) describe a statistical approach in which large quantities of human–human dialogue was categorized into dialogue acts. Their model uses lexical, prosodic and collocational cues, and also includes aspects such as discourse coherence.

6.4 Ethical Considerations

When pursuing empirical human–computer studies with human subjects, it is important to keep in mind that the researcher has ethical obligations to each person participating in an experiment. In studies where speech data is collected and analyzed, these questions become critical, since every person's voice is a unique and personal imprint. According to *Vetenskapsrådet* (the Swedish Research Council), four general ethical principles must be considered during all experiments which involve data collection with human subjects (HSFR 1994). These guidelines are listed below, along with some implications for each principle:

- The principle of information
  - Subjects should be informed about the purpose of the research they are contributing to.
- The principle of consent
  - Potential participants in a research project have the right to decide if they want to be in it. They should be informed that they are under no obligation to participate, and that they are free to terminate their interaction at any time. For subjects who are under the age of 15, parental consent is required.
- The principle of confidentiality
  - All personal data collected about individual subjects should be kept confidential, and should be stored as to avoid unauthorized access.
- The principle of usage
  - Information about individual subjects may only be used for research purposes, and cannot be used for commercial or other purposes.

During data collection in a controlled laboratory environment, it is relatively easy to create a protocol by which the above principles are followed, as discussed in Section 6.2.1. However, when data is collected with an unsupervised system in a public environment, it can become more difficult to adhere to the principles.

6.5 Summary and Implications

In this chapter, we have argued for using empirical methods in human–computer interaction research. Human behavior in general is highly diverse and displays great individual variability. Accordingly, introspection is insufficient and empirical data collection is necessary when designing natural language interfaces. However, there are problems associated with the empirical approach and the design of speech interfaces. For instance, the freedom that natural language interfaces allow their users can make it difficult for them to know what to say. An iterative design process, where a system is repeatedly tested and improved, is one way of dealing with this problem.

In the second part of the chapter, we discussed different methods of data collection in the development of spoken dialogue systems. Data collection using simulated Wizard-of-Oz systems was compared to using fully functional systems. Similarly, arguments for and against using controlled and unsupervised data collection methods were weighed against one another. In brief, it was concluded that both types of methods and both types of systems have their individual merits. In simulated experiments in controlled settings, factors related to the user interface and dialogue context can mostly be kept static. Individual differences in user behavior in corpora collected with such systems can to a large extent be ascribed to background factors, for instance the user's prior expectations. On the other hand, fully functional systems elicit more realistic errors and the absence of anyone surveying the user's behavior should make the human–computer dialogues less constrained.

Next, we discussed the annotation of spoken dialogue corpora. We noted that the analysis of speech data requires several additional levels of transcription and labelling when compared to text corpora. The assignment of speech act labels to user utterances was also briefly discussed.

Finally, the chapter was concluded with a brief discussion about ethics. Data collection with spoken language interfaces makes certain ethical considerations necessary, since it involves the recording and analysis of individual and personal information.

The second part of this thesis, which begins with Chapter 7, describes four spoken dialogue systems and corpora. The human–computer dialogue corpora were collected using both Wizard-of-Oz systems in laboratory environments as well as unsupervised fully functional systems in public settings. The analyses of the four corpora, presented as nine empirical studies in Part II, indicate that the different conditions under which the human–computer interactions took place affected user strategies. In particular, the design of the user interface and the setting of the data collection were factors which contributed to the way users adapted their language in these systems.

Part II

Chapter 7

Four Spoken Dialogue Corpora

This chapter deals with the process of collecting the spoken dialogue data that is the focus of study in this thesis. The spoken dialogue systems used to collect the four corpora are briefly described, after which the specific details of each data collection which may have contributed to the outcome of the dialogues are discussed.[2] Each section contains an example dialogue, a picture of the user interface and some corpus statistics. Detailed descriptions of the analyses of corpora are presented in the respective subsections of Chapter 7. In the last section of the chapter, we present a brief overview of the included corpora in the form of a table, highlighting relevant properties. In addition, we discuss some details of the data collection that may have contributed to the outcome of the dialogues, and some lessons learned in the process of collecting the spoken dialogue corpora.

7.1 August

The August project was initiated and carried out at the Centre for Speech Technology at KTH during 1998 and 1999. August is a Swedish multimodal spoken dialogue system featuring an animated agent (Gustafson, Lindberg and Lundeberg 1999; Lundeberg and Beskow 1999). August was named after the Swedish turn-of-the-century author August Strindberg, and was exhibited as part of an effort to promote KTH in general and speech technology in particular. The system was put on display downtown at Kulturhuset (the Cultural Centre) as part of the event 'Stockholm Cultural Capital of Europe 1998.'

---

[2] For detailed descriptions of the August system, the AdApt system and the Pixie system, see Gustafson (2002).

Figure 7.1 The set-up and graphical user interface of the August system.

7.1.1 System Description

Instead of presenting users with a single complex domain, the system was designed to enable interaction in a number of shallow domains. August provided users with general information about the Royal Institute of Technology (KTH), speech technology and Stockholm. By extracting information from web-based yellow pages, August also helped users with information about the location of restaurants and other facilities in the city. Moreover, August had access to some basic facts about the life and works of his namesake Strindberg. Finally, because the users did not always know what to say to August and since his appearance was rather human-like, it seemed likely that they would want to exchange greetings with him. The ability to handle and respond to some of these social utterances was also built into the system.

When developing the August system, the Wizard-of-Oz approach was not used. Instead, this experimental system was exposed to real users from the general public from the beginning. The users of the system were not given any explicit instructions. To suggest possible topics of conversation when no one was interacting with August, facts related to the domains of the system were randomly selected and read out loud by the animated agent. The agent also had a 'thought balloon', where domain-related ideas appeared. The system's responses differed both in length and complexity, from simple single-word utterances to long phrases. As long as a user asked something within one of its domains, the system often gave a reasonable response. This resulted in a system that sometimes appeared to handle almost anything and which managed to generate quite humanlike responses, while it sometimes did not seem to 'understand' much at all. An example dialogue can be seen in Table 7.1.

The system was put together in the first half of 1998. From August 1998 to March 1999, the spoken dialogue system was available six days a week to all visitors at the Cultural Centre in downtown Stockholm. August was set up in a public exhibition area with high levels of background noise from other equipment and visitors, which made it necessary to install a click-to-talk mechanism for recording the speech input. Since the system was unsupervised and the equipment had to be protected, a directional microphone secured in a metal grid box was installed. The metal box introduced some sound deterioration but experiments indicated that this would not effect the speech recognition significantly (Gustafson, Lundeberg and Liljencrants 1999). The system had two computer screens, as can be seen in Figure 7.1. The first screen featured the animated agent who communicated with synthetic speech and a combination of facial expressions and head movements. The second screen was used for presenting tables and a street map, where locations of items that matched the users' requests were shown.

A new lip-synchronized 3D talking head was developed for the August project (Lundeberg and Beskow 1999; Beskow 2003). To give the agent a 'personality,' he was made to resemble the author August Strindberg. Strindberg is a famous literary character, and the choice of this persona also indicated that the system had some knowledge about Stockholm, history and literature. When designing the agent, it was important that he should not only be able to generate convincing lip-synchronized speech, but that he would also have access to a repertoire of non-verbal behavior. Facial movements which displayed emotions and directed the user's attention to physical objects in the environment were developed. Furthermore, the synthetic speech output was accentuated using gestures to stress certain focussed words and phrases. Speech synthesis parameter trajectories were generated by the KTH audio-visual text-to-speech system (Carlson, Granström and Hunnicutt 1982). The speech output was generated using an Mbrola synthesizer (Dutoit et al. 1996), with a Swedish voice created by Gösta Bruce and Marcus Filipsson at Lund University.

One important issue in the August project was that of extending the coverage of user utterances, so that the system could be improved by adding a new domain, or altering an existing one. After the August system had been on display for three months, all user utterances recorded so far were transcribed and analyzed. The system was then updated in accordance with the most frequently observed user behavior. In this process, the recognition lexicon, the semantic analyzer and the system output was modified. Furthermore, the analyses implied that the users sometimes had difficulties understanding that the system was processing their input. To enhance the perceived reactivity of the animated agent, listening and thinking gestures were created.

Table 7.1 Example dialogue from the August corpus.

User	August
Berätta någonting *Tell me something*	Hej, jag kommer från institutionen för tal musik och hörsel *Hello, I come from the department of speech, music and hearing*
Vem är du? *Who are you?*	Jag är en djevla man, som kan göra många konster *I am one hell of a man, and I can do a lot of tricks*
Berätta mera *Tell me more*	Men jag har snygga mustacher *But I have great looking moustaches*
Var bor du? *Where do you live?*	Jag kan till exempel berätta om Stockholm, KTH och August Strindberg *I can for example tell you about Stockholm, KTH and August Strindberg*
Berätta om August Strindberg *Tell me about August Strindberg*	Intima teatern startades av honom *The Intima theater was founded by him*
Vilken teater? *Which theater?*	Sådana frågor får ni ställa till SLs trafikupplysning *Such question will have to be answered by the local transportation information office*

7.1.2 The August Speech Corpus

In this section, the August corpus and some of the linguistic analyses of this database are presented. The August corpus has previously been described in a number of papers (Bell and Gustafson 1999a; Bell and Gustafson 1999b; Gustafson and Bell 2000).

After cleaning the data by removing empty soundfiles, the August corpus was found to consist of 10,058 spontaneous utterances. All of these utterances were transcribed orthographically, part-of-speech tagged, parsed and labeled with some basic speaker and utterance characteristics. The 39,594 word tokens in the corpus made up a lexicon of 2,915 word forms, out of which half occurred only once. The 200 most frequently used word forms covered 81% of all words in the database. Most of the utterances consisted of a single clause, and coordinated or subordinated clauses rarely occurred. The average number of words per utterance was about four, which partly explains the relatively small number of syntactically complex structures in the corpus. The syntactic parsing resulted in phrase level sentence structures, for example (NP) (VP) (NP) (PP). More than half of all the utterances in the database could be covered by ten such structures.

August was designed as an unsupervised, stand-alone spoken dialogue system, and the original database consists of sound files alone. Background noise from other visitors made some of the utterances difficult to transcribe. Furthermore, due to the lack of background information regarding the speakers, the categorization of the data was not unproblematic. For example, it was problematic for the August transcribers to know when the same user continued to speak and when a new user appeared. In addition, it was at times hard to assess whether a certain speaker was a man, woman or child. We also noticed that some speakers came in groups and took turns at speaking to the system. Because the system remained in use for such an extended period of time, it was virtually impossible to establish whether an individual user talked to August for the first time or had come to revisit the system. The figures given for distribution and number of speakers in the corpus must be seen as rough approximations. The number of speakers in the August corpus was estimated at 2,685, out of which about one half were men, and one quarter women and children, respectively. The number of utterances originating from a single speaker in the corpus ranged from one to forty-nine, but the average was 4.1 utterances for men, 3.3 for women and 3.5 for children. The corpus statistics are summarized in Figure 7.2.

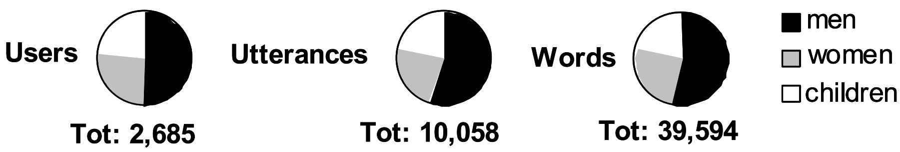


Figure 7.2 The distribution of users, utterances and words in the August corpus.

7.2 AdApt

The AdApt system was developed at the Centre for Speech Technology at KTH between 1999 and 2000 (Gustafson et al. 2000). The system was developed by collecting data in a Wizard-of-Oz experiment (Bell, Boye, Gustafson and Wirén 2000), which was subsequently analyzed and used to implement and evaluate a fully functional version of the system in 2001-2002 (Bell, Boye and Gustafson 2001; Edlund and Nordstrand 2002). The goal of the AdApt project, which was developed jointly by researchers at KTH-CTT and Telia Research, was to explore the possibilities of a conversational spoken dialogue system that could handle multimodal input as well as produce multimodal output.

7.2.1 System Description

The domain of the AdApt system was chosen to be the apartment realty market of downtown Stockholm, with the animated agent Urban as a virtual real-estate agent. Part of the motivation behind selecting this particular domain was the fact that many people are interested in apartments, and at least occasionally want to acquire a new one. Moreover, the domain appeared to be advantageous from the point of view of examining multimodal input behavior and designing appropriate multimodal system output. Apartments are complex objects with features that are suitable for graphical presentation (location in the city), as well as properties suitable for verbal presentation (price, description of interior details). Furthermore, it is not always obvious which modality is preferable for a referential construction. Users of the AdApt system could refer to apartments they were interested in verbally, graphically or by combining the two modalities e.g. by saying "this one" while clicking on an object on the map. The system's graphical user interface, with the animated agent Urban and the interactive maps, can be seen in Figure 7.3.

The technique used for the animated agent Urban in the AdApt system is based on the same system as the face developed for August, see Section 7.1. Both systems employ a parameterized, deformable 3D facial model controlled by rules within a text-to-speech framework. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account (Beskow 2003).

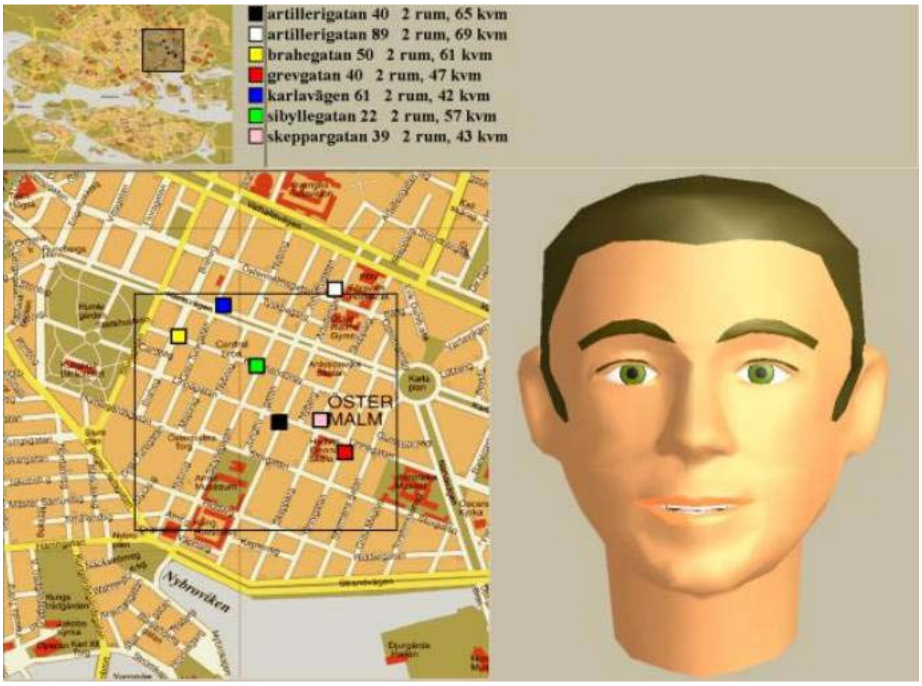


Figure 7.3 The AdApt user interface.

7.2.2 The AdApt Corpus

Data collection for a Wizard-of-Oz experiment took place in the spring of 2000. For the purpose of this data collection, the system's key functionalities (analysis of multimodal user input, dialogue management and multimodal response generation) were handled by a human operator (the wizard). The simulated system's verbal and graphical output was generated by means of predetermined templates. However, the wizard's turn-taking behavior or dialogue management capabilities were not subject to the same restrictions. Subjects experienced a system with an animated agent who was somewhat rigid in appearance and whose repertoire of natural language output was limited. On the other hand, the animated agent Urban exhibited advanced understanding skills as well as discourse abilities. An example of a dialogue between a user and the AdApt system can be seen in Table 7.2.

Table 7.2 Example dialogue from the AdApt WOZ corpus.

System	**User**
Jag heter Urban och jag kan berätta om bostäder i Stockholm *My name is Urban and I can tell you about apartments in Stockholm*	Visa lägenheter i Vasastaden och Östermalm *Show apartments in Vasastaden and Östermalm*
Hur många rum vill du ha? *How many rooms do you want?*	Ett till två *One to two*
Jag hittar för många lägenheter för att visa . I vilket område ska jag leta i första hand? *I find too many apartments to show. In which area should I begin looking?*	Vasastan *Vasastan*
Ett område på kartan är markerat. Där finns det just nu 8 lägenheter till salu *An area on the map has been marked. There you can find 8 apartments for sale*	Visa lägenheter byggda mellan 1900 och 1950 *Show apartments built between 1900 and 1950*
I det markerade området hittar jag 6 lägenheter *In the marked area there are 6 apartments*	Eh berätta om Frejgatan 51 *Eh tell me about Frejgatan 51*
Den svarta huset är byggt 1926 *The black building was constructed in 1926*	Vad finns i lägenheten? *What is in the apartment?*
Den svarta fastigheten har kabeltv *The black building has cable tv*	Trevligt finns det balkong *Nice is there a balcony*
Nej *No*	Visa lägenheter ur samma urval med balkong *Show apartments from the same selection with a balcony*
Hittar inga sådana *Cannot find any*	Okej utöka sökning till ett större område *Okay extend the search to a larger area*

A complete description of the experimental setup can be found in Bell, Boye, Gustafson and Wirén (2000). In brief, the purpose of the Wizard-of-Oz experiment was to examine whether it was possible to influence the user's choice of modality by suggesting different ways of referring to apartments. The system either referred to apartments verbally by using color codes corresponding to icons on the map ("the red apartment"), or graphically by highlighting a particular icon and saying "this one." The data collection took place at the Centre for Speech Technology, where subjects were asked by the experimental leader to begin by providing some background information regarding previous experience with computer interfaces in general and speech technology in particular. Subjects were then informed about the overall idea behind the research and setup of the experiment, and commenced their interaction. After finishing their interaction with the AdApt system, each subject was asked to provide the experimental leader with feedback on the system's functionality. They were then informed that the system had been a simulation, and the experimental leader made sure that this came as a surprise to all of the subjects involved. Each dialogue session lasted for about 30 minutes.

The AdApt corpus consists of 32 dialogues with 16 subjects, out of which eight were men and eight women. The subjects' ages ranged from 17 to 55. A few of the subjects were employees of the Department of Speech, Music and Hearing, but none of them involved in speech technology research. The other subjects were recruited from outside the department. The total number of utterances in the AdApt corpus is 847. On average, an utterance in the AdApt database consisted of 7.6 words, and a single utterance varied in length between one and 47 words.

All utterances in the database were orthographically transcribed, tagged, parsed and labeled for disfluencies. In the analysis of the corpus, many utterances were found to consist of fragments separated by silent pauses. In some cases, the fragments could be interpreted as complete but in other cases more input was needed to make sense of the utterance. Another finding in the AdApt corpus was that some subjects consistently used positive and negative feedback to signal understanding and non-understanding throughout the dialogues. Along with the occurrence of disfluencies in the corpus, these phenomena became the subject of separate empirical studies described in the sections below.

7.3 Pixie

The Pixie system (Gustafson and Sjölander 2002) was developed at Telia Research as part of the exhibition 'Tänk Om' ('What If'), which opened in December 2001 at the Telecom museum in Stockholm. The system's speech interface was developed at Telia Research, while the system's graphical user interface and animated agent were developed by Liquid Media, a computer games company in Stockholm.

7.3.1 System Description

At the ongoing exhibition at the Telecom museum, visitors can experience a full-sized 'intelligent' apartment of the year 2010. The apartment contains conceptual prototypes of future household innovations, such as an interactive wall for communicating and a robot pet. Apart from displaying these objects, the apartment is intended to make visitors consider how innovations in the areas of interface design and communication may come to change the way we relate to technology in the near future. The animated agent Pixie, (see Figure 7.) with whom the users interact in spoken Swedish, is supposed to visualize an embodied speech interface to both information services and home control in this apartment.

Figure 7.4 The Pixie interface and a family interacting with the system.

Members of the general public who visit the museum enter the 'Tänk Om' exhibition in groups of up to 25. Before entering, they must register at a computer by entering some personal background information such as age and gender. At this time, users are informed that their participation is voluntary but if they do choose to register their speech will be recorded and used for research purposes. Those visitors who continue get their personal login information stored in the system's database and simultaneously encoded into a smart card. To begin with, all visitors watch a film which introduces the apartment and Pixie. This film gives the users some suggestions about which areas of conversation Pixie might be able to engage in. Next, visitors enter the apartment where twelve computer screens have been built into walls and tables. By inserting their individual smart card into the slot next to the screen, visitors are able to engage in spontaneous dialogue with Pixie by talking into handheld microphones. Graphical icons on the screen indicate when it is possible to talk to Pixie, and when she is 'thinking' about the user's previous utterance. The visitors are told they could either help Pixie perform certain tasks in the apartment or ask her general questions about herself or the exhibition. An example dialogue of the latter category can be seen in Table 7.3 below.

Table 7.3 User interacting with the Pixie system.

User	**Pixie**
Kan du laga mat? *Do you know how to cook?*	Du ser ut som det godaste härinne. Nu kommer jag och äter upp dig, hi hi! *You look like the tastiest thing in here. Now I'm coming to gobble you up, hee hee!*
Vad vet du om mig? *What do you know about me?*	Du är bäst, inte sant? *You are the best, right?*
Hur gammal är jag? *How old am I?*	Ursäkta, nu hörde jag inte vad du sa *Sorry, I didn't quite catch that*
Hur gammal är jag? *How old am I?*	Låt mig tänka, när aktiverades jag, det måste ha varit i december 2001, men jag tror jag är lite smartare än en nyfödd *Let me think, when was I activated, it must have been in December 2001, but I think I am a bit smarter than a newborn baby*
Vad heter jag? *What is my name?*	Du är bäst, inte sant? *You are the best, right?*
Har du någon familj? *Do you have a family?*	Jag skapades av ett gäng forskare från Telia Research, men min familj är förstås familjen Ström *I was created by a bunch of researchers at Telia Research, but my family are the Ströms of course*

7.3.2 The Pixie Corpus

After the first year of recordings, more than 100,000 utterances of spontaneous computer-directed Swedish speech have been collected. Out of these, 25,000 utterances recorded during the first six months were manually checked and transcribed. Empty files containing only background noise were removed and utterances were orthographically transcribed and labeled with tags for exaggerated pronunciation in terms of loudness and hyperarticulation. In the transcribed part of the database the number of speakers is 2,885, which means that the average number of utterances per speaker is about nine. Because of the speaker pre-registration and the usage of smart cards, it was possible to get a more accurate picture of the users of this system than was the case with the August system, see Section 7.1.

The Pixie database contains information about the gender and age of each speaker, which in the case of children speakers can be very difficult to assess by only listening to a recording. This information is valuable both for research purposes and the improvement of acoustic models by training the speech recognizer on children's speech (Gustafson and Sjölander 2002). On the whole, the solution with user pre-registration and smart cards turned out to be very reliable. However, once inside the exhibition there was no one to check that users did not exchange smart cards with one another. During labelling, the transcribers had access to the database entry field e.g. "John Anderson/ Male/ age: 57." If, for instance, the voice of a child could be heard when the above card was used, the transcribers of the Pixie database could infer that someone other than the card holder had taken over the microphone. These utterances were tagged with "wrong speaker," and were subsequently removed from the database.

7.4 NICE

NICE is a three-year project within the European Union's Human Language Technologies (HLT) program. The goal of the project is to develop a speech-enabled computer game which allows both children and adults to engage in natural and fun communication with embodied literary characters using several modalities. The project was initiated in 2001 and was officially launched in March, 2002. Apart from Telia Research, the following partners are members of the project: NISLab (University of Southern Denmark), Philips speech processing (now ScanSoft, Aachen), Liquid Media, Stockholm and LIMSI (Centre National de la Recherche Scientifique, Paris).

7.4.1 System Description

The aim of NICE is to create a multimodal interactive dialogue system by developing a conversational interface for communication between humans and embodied historical and literary characters from the fairy tale universe of H C Andersen. One of the emphasized research issues in the project is how to develop interactive systems that are suitable and enjoyable for children and adolescent users. In particular, the project will focus on improving the interpretation of children's speech and natural language input. However, for practical purposes the first data collection was aimed at adult users.

The first version of the system was a Wizard-of-Oz simulation. One of the co-authors of the study (Bell, Gustafson and Heldner, 2003), sitting behind a screen, acted as the system's speech recognition and dialogue management modules following a template. Subjects were randomly assigned to interact with either one of two versions of the simulated system: one version in which the speaking rate of the feedback prompts was increased, and one in which it was slowed down. All other aspects of 'fast' and 'slow' Cloddy Hans were kept identical. Cloddy Hans' spoken output was generated in the following way: One of the researchers involved in the experiment recorded the prompts, enacting Cloddy Hans' personality and speaking style. The speech was then manipulated to simulate vocal tract lengthening and lowering of f_0, using a TD-Psola algorithm (Gustafson and Sjölander 2002). The same algorithm was also used to produce the two test sets by increasing or decreasing the speech rate of the original utterances by 30%.

The study was performed in the exhibition area of the Telecom museum in Stockholm. Users sat on a couch in the livingroom of the apartment described in Section 7.3.1, and interacted with a full-sized animated character, Cloddy Hans. Figure 7.5 depicts one of the users interacting with Cloddy Hans.

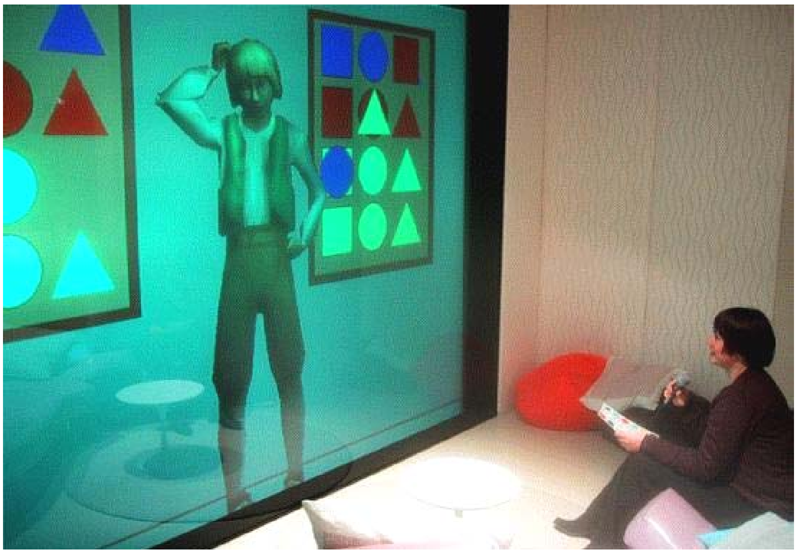


Figure 7.5 A user interacting with the Cloddy WOZ set-up.

7.4.2 The Cloddy Hans Corpus

In the spring of 2003, a first Wizard-of-Oz experiment with one of the full-body fairy-tale characters was set up at the Stockholm Telecom museum. The hypothesis behind this experiment, described in detail in Bell, Gustafson and Heldner (2003) as well as in Chapter 8 below, was that users of spoken dialogue systems adapt their speech rate to that of the system's output.

In the course of the experiment, sixteen volunteer subjects were recruited. The subjects were 9 men and 7 women between the ages of 17 and 59. Most of the subjects were members of the general public, and some were museum employees. Before agreeing to be part of the experiment, the subjects were informed about the general purpose of the study and were told that they were being recorded. Subsequently, subjects were given a pictorial scenario and instructions on how to talk into the microphone, and were told to await further instructions. Before starting their actual interaction with the system, the system was calibrated by asking subjects to read five sequences of digits. After each of these sequences, the system echoed the digits. The subjects then listened to a prerecorded instruction on how to proceed with the actual task. This involved helping the fairy-tale character Cloddy Hans to solve a puzzle. Subjects were told that the solution to the puzzle would be revealed to them after a number of colored geometrical figures had been moved from one part of the screen to another in a certain order. Since Cloddy Hans lacked information required to perform this task, the subjects had to help him using their pictorial scenarios.

The total number of recorded user utterances was 297. Each dialogue consisted of 6 tasks (i.e. there were 6 colored geometrical figures to be moved) and 16 to 27 user turns. The total number of words in the corpus was 2,173. The entire corpus was orthographically transcribed, and all user turns were tagged with information on position in the dialogue, type of user turn, previous system output, etc. At the word level, the user utterances were also labeled for lexical content, distinguishing between *color*, *shape* and *other* words. All utterances were also labeled with tags for when speakers had adapted their syntax and/or lexical choices to match that of the system. The users' way of expressing themselves throughout the dialogues could thus be analyzed in detail. An example dialogue can be seen in Table 7.4.

Table 7.4 The first part of a dialogue from the Cloddy Hans corpus.

User	System
Hej ta den röda fyrkanten och sätt på den blåa trekanten *Hello take the red square and put it on the blue trigon*	Vill du att jag ska flytta den röda fyrkanten till den blåa trekanten? *Do you want me to move the red square to the blue trigon?*
Ja det stämmer *Yes that's right*	Ok *Ok*
Bra ta nu den blåa cirkeln och lägg på den gröna fyrkanten *Good now take the blue circle and lay it on the green square*	Det där förstod jag inte *Now I didn't understand*
Den blåa bollen på den på den gröna fyrkanten *The blue ball on the green square*	Det där förstod jag inte *Now I didn't understand*
Raden högst upp den blåa runda bollen i mitten läggs på den gröna fyrkanten *The row on top the blue round ball in the middle is lain on the green square*	Vill du att jag ska flytta den blåa cirkeln till gröna fyrkanten? *Do you want me to move the blue circle to the green square?*
Ja det stämmer *Yes, that's right*	Ok *Ok*
Bra ta nu den gröna triangeln och sätt den på den röda cirkeln *Good now take the green triangle and put it on the red circle*	Vill du att jag ska flytta den gröna trekanten till den röda cirkeln? *Do you want me to move the green trigon to the red circle?*
Det stämmer *That's right*	Ok *Ok*
Bra nu tar du den gula cirkeln och lägg på den gröna triangeln *Good now you take the yellow circle and lay it on the green trigon*	Vill du att jag ska flytta den blåa fyrkanten till den röda cirkeln *Do you want me to move the blue square to the red circle?*
Nej den gula cirkeln på den gröna triangeln *No the yellow circle on the green triangle*	Vill du att jag ska flytta den blåa cirkeln till den gröna trekanten *Do you want me to move the blue circle to the green trigon?*

7.5 Corpora Overview and Discussion

In this section, we present an overview of the system and corpora described in Chapter 7. In addition, we address some of the factors which may have influenced the results of the data collection and discuss some things we learned in the process. A more detailed description of the first three corpora can be found in Gustafson (2002).

As can be seen in Table 7.5 below , the included four corpora were collected under diverse conditions, and vary considerably in terms of size and number of speakers:

Table 7.5 An overview of the included corpora.

System	Place	Users	Utterances	Type of system	System introduction
August	Public exhibition place	2,685 visitors general public	10,058	Fully functional	Indirect
AdApt	Lab at KTH	16 invited subjects	866[3]	WoZ	Verbal from experiment leader
Pixie	Museum	2,885 visitors	25,000	Fully functional	Movie museum guide
NICE	Museum	16 visitors	297	WoZ	Verbal from system itself

In the course of collecting the four dialogue corpora examined in the current thesis, the advantages and disadvantages of collecting data with both fully functional as well as Wizard-of-Oz systems were experienced hands-on. Two of the corpora, August and Pixie, were collected by using stand-alone fully functional systems which were exposed to members of the general public in their respective locations. As described above, the August system was totally unsupervised, while the Pixie dialogues were collected in a museum with guides

[3] In the AdApt database, the number of user turns is 866. However, because we used an open microphone and the wizard of the system waited until he regarded the user's input as 'complete', a turn is sometimes made up of several utterances. Further details about the users' turn-taking behavior in the AdApt corpus can be found in Section 9.2.

present in the exhibition and registration areas. The other two corpora, AdApt and NICE, were collected with Wizard-of-Oz systems in which a human operator simulated parts of the system's understanding capabilities. While the AdApt corpus was collected at a research lab with invited subjects who came to be part of an experiment, the NICE dialogues were collected at a museum where random visitors were asked to contribute to a research project by interacting with a spoken dialogue system.

As discussed in Chapter 6, one disadvantage of collecting data with a stand-alone system is that it is difficult to maintain control of the speakers of such a system. For a given utterance, it can be difficult or even impossible for the transcriber of the corpus to know when the current speaker continues the dialogue and when someone new starts speaking with the system. At worst, this can make analyses at the dialogue level methodologically unsound. What *is* a 'dialogue' when you can only be certain about the identity of one of the speakers? In a system that you can simply walk up to, it can also be difficult to assess whether an individual speaker is a man, woman or child. This makes it difficult to evaluate whether children's speech was different than adults'. As a result of the experiences learned in the development of the August system, we designed the environment of the Pixie data collection differently. The smartcard solution, described in Section 7.3, made it possible to get a good estimate of who was speaking to the system when, and for how long. On the other hand, the Pixie dialogues were perhaps not quite as unconstrained as the August dialogues. During the users' interaction with Pixie, guides were present in the room. Although the guides' behavior was non-intrusive, some users may have felt a need to restrain themselves.

The AdApt corpus was collected by inviting subjects to the lab, and carefully introducing them to interacting with the system. To avoid giving the subjects different instructions, the experimental leader read the subjects' background information out loud from a script. The task was presented in the form of a pictorial scenario, which prevented lexically priming the users. Despite these precautions, post-experimental interviews indicated that some of the subjects felt influenced by the fact that they were visitors at a research laboratory, and that their behavior may have been affected. One of the research questions addressed in the wizard simulation of the AdApt system was to what extent users of a multimodal system would be influenced by the choice of modality on part of the system (graphic vs. verbal). The results showed that many of the subjects seemed to have an overall preference to speaking when compared to clicking and drawing. When interviewed about their modality preferences after their session with the system, one of the subjects said "I thought I was supposed to use my voice. After all, I was asked to be part of a research experiment at the Centre for Speech Technology." Because of these experiences, the NICE corpus was collected in a semi-public environment, where random visitors at the Telecom museum were asked if they would consider participating in a research experiment. In this way, we could ensure that the participants felt no obligations towards the researcher and her goals. Subjects were asked to sit down in front of a big screen, and a prerecorded voice instructed them on what to do. By that time, the experimental leader had left the room and subjects had no one to ask if they were uncertain about what to do.

After this overview of the four dialogue systems and data collections, it is now time to present the findings of the resulting spoken language corpora. Part II deals with user adaptation at different linguistic levels, and each of the chapters includes several empirical investigations.

Chapter 8

Phonetic and Prosodic Adaptations

As discussed in Chapters 4 and 5, people often adapt their manner of speaking during human–computer interaction. Moreover, patterns of adaptive behavior are sometimes magnified if the dialogue becomes problematic. In this section, three empirical studies dealing with phonetic and prosodic aspects of user adaptations are presented. The first and second studies both focus on how people who interact with a spoken dialogue system in a public location respond to errors and miscommunication by modifying their speech during repetitive sequences. The third study investigates if people adapt their speech to the speaking rate of the system they are interacting with.

In the first study, which is a revised version of an article written with Joakim Gustafson for *ICPhS* (International Congress of Phonetic Sciences), original input to the August system is compared to lexically identical repetitions of these utterances (Bell and Gustafson 1999). Joakim Gustafson did most of the work of designing and implementing the August system, and was also responsible for the task of supporting and maintaining the system during the six months when it was used to collect data. Several other people who contributed to the project are acknowledged in Gustafson, Lindberg and Lundeberg (1999). Linda Bell made some contributions to the design and development of the system during the latter part of the project. Both authors contributed equally to all aspects of tagging and labelling the corpus, as well as analyzing the results.

The second study, based on an investigation performed with Joakim Gustafson and previously published in the *Proceedings of Eurospeech*, focuses on phonetic analyses of longer sequences of identical and non-identical repetitions in the Pixie corpus (Bell and Gustafson 2003). While both authors were involved in the development of the Pixie system, Joakim Gustafson also made a substantial contribution to the implementation of this system. Linda Bell and

Joakim Gustafson both contributed equally to all aspects of tagging and labelling the corpus, as well as analyzing the results.

The third and final study is a revised version of a paper written with Joakim Gustafson and Mattias Heldner. The results from the experiment were first published in the *Proceedings of ICPhS* (Bell, Gustafson and Heldner 2003). The design and development of the system setup, as well as the data collection, was done by the authors in collaboration. Joakim Gustafson was responsible for the implementation of the wizard environment. Mattias Heldner developed the technique for measuring speech rate, and did the statistical analyses.

8.1 Repetitive Sequences in the August Corpus

In this section, sequences of original utterances and identical repetitions in the August database are investigated. To examine whether repetitions could be distinguished from original input, acoustic-phonetic analyses of the sound files were performed.

8.1.1 Background

When a listener fails to understand what someone says, she often asks the speaker to repeat the previous utterance. Although repetition can have several different functions in human–human dialogue (Swerts et al. 1998), the pattern of listener non-understanding followed by a speaker verbatim or near-verbatim repetition is well-established in spontaneous conversation. In the context of spoken dialogue systems, speakers often repeat themselves when they believe that their original input to the system has been ignored or incorrectly handled by the system. These repetitions have been reported to display certain acoustic-phonetic features. Previous studies have shown that computer-directed repetitions are characterized by a greater number of pauses between words, a movement towards clearer articulation, a suppression of disfluencies and an increase in utterance duration (Levow 1998; Oviatt et al. 1998c).

For users of unimodal spoken dialogue systems who wish to resolve errors in the dialogue, repetition is one of few strategies that are easily available. Studies of multimodal systems have shown that users often switch to an alternative mode if their spoken input is not correctly handled (Oviatt et al. 1998a). However, the August system, described in 7.1, had no optional input channel. Instead, users could only try to increase the linguistic contrast in the single channel that they had access to.

8.1.2 Data and Annotation

For the purpose of the present investigation, the first half of the August corpus was examined in detail. 4,647 utterances spoken by approximately 1,380 users were extracted with the purpose of studying repetitive sequences. In this part of the corpus, repetitions of all kinds constituted about 10% of all utterances. Half of these were lexically identical repetitions, which for the purpose of the current phonetic analysis were considered the most interesting.

452 utterances (200 original utterances and 252 repetitions) were thus extracted from the August database and manually labeled. About one fourth of these utterances were judged to be spoken by children. Pairs consisting of an original utterance followed by a single repetition were the most frequently occurring pattern in this sub-section of the corpus, but an original input utterance was repeated verbatim up to five times in succession. Figure 8.1 is example of a repetitive sequence with an original utterance followed by two verbatim repetitions.

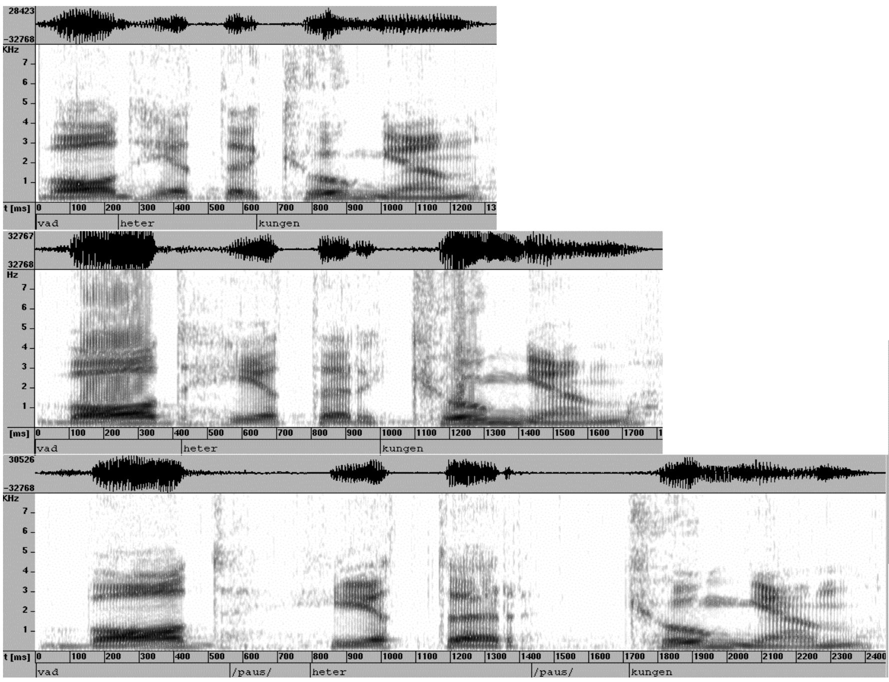


Figure 8.1 The utterance "Vad heter kungen?" ("What is the name of the king?") as original input (top) and repeated twice by the same speaker.

In the comparison between original input and repetitions, the following features were the ones examined:

- utterance duration
- speech rate
- degree of hyperarticulation
- focus (stress)
- increased loudness and
- inserted pauses

The corpus was coded and labeled partly by hand, with the present authors perceptually assessing the features *degree of hyperarticulation*, *stress* and *loudness*. Because the users of this unsupervised system spoke at various distances from the microphone, it was not possible to use an objective measure of *loudness*. *Utterance duration* was computed by removing any silences at the beginning and end of each sound file, after which a comparison of the original input and corresponding repetition(s) could be made. *Speech rate* was approximated using the measure of number of syllables per second.

8.1.3 Results

Repetitions were found to be on average 15% longer than the original utterances. This figure is comparable to previously reported findings for verbatim repetitions in computer-directed English (Oviatt et al. 1996; Levow 1998). For adults, the difference in utterance duration was 18%, while children's repetitions were found to be on average 7% longer than their original input to the August system. Although most users spoke slower during repetition, this was not always the case. A few speakers did not increase the duration, and some even spoke faster during repetition. This makes the average figures reported above somewhat misleading in the sense that they include both increases and decreases in utterance duration. Our analyses showed that the average lengthening of duration was about 40%, and the corresponding figure for shortening of duration was 15%. The distribution of durational changes in repetitions is illustrated in the histogram in Figure 8.2.

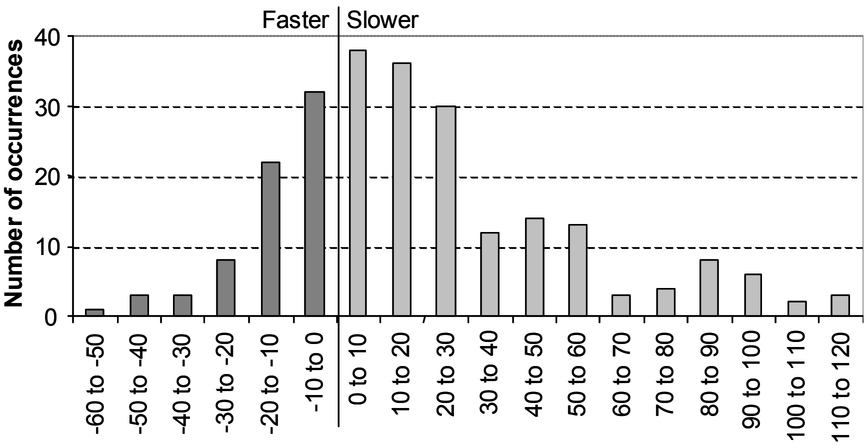


Figure 8.2 Distribution of durational changes in repetitions when compared to original utterances. Each bar represents the number of repetitions, with durational changes in ten-percent steps.

In the cases where the same utterance was repeated more than once, over half of the second repetitions were found to be longer than the first repetition. The second repetitions were also distinguished by the fact that they more often contained inserted pauses between words. Inserted pauses were found in 29% of the second repetitions compared to 7.5% of the first repetitions and 2.5% of the original utterances. This can be seen in the example in Figure 8.1., in which the user's second repetition contains inserted pauses between words.

A decrease in speech rate during repetition was found mainly in those cases where the user spoke normal or fast to begin with. For users whose original utterance was labeled as slow in speech rate, no adaptation appeared to take place. It is possible that because of their initial expectations, these users had already in a sense 'adapted' themselves to the system by speaking slower.

As can be seen in Figure 8.3, 40% of the adults' and 28% of the children's repetitions were labeled as more clearly articulated than the original input to the system. However, a small number of repetitions were labeled as less clearly articulated than the original utterances. It has previously been reported that while loudness is increased during error resolution in human–human communication, this is not the case when people speak to computers (Oviatt et al. 1996). In the August corpus, 21% of children's repetitions were found to be increased in loudness. A corresponding tendency was not found in the adult user group, where merely 5% of the utterances had been labeled as increased in loudness. Finally, Figure 8.3 also shows that shifting of focus occurred in 17% of the adult repetitions and in 7% of the childrens'.

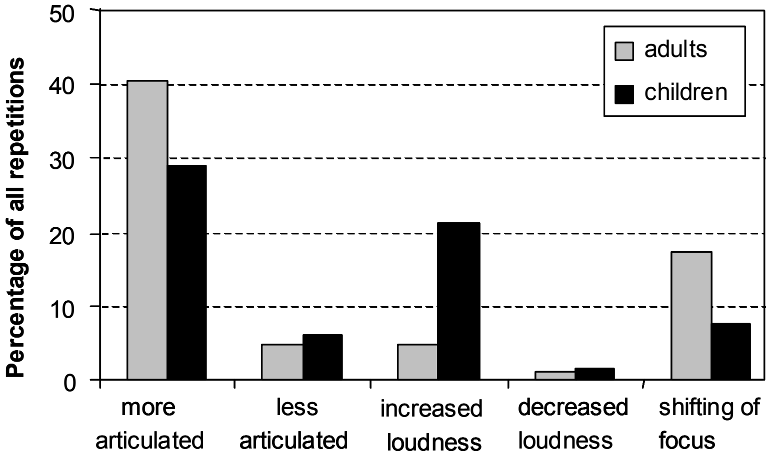


Figure 8.3 Distinguishing features in repetition.

8.1.4 Discussion and Summary

Most people adapt their manner of speaking to meet what they believe to be the demands of a spoken dialogue system. One third of the verbatim repetitions in the current corpus, however, were not labeled as different from the original input to the system. Little or no adaptation took place, which could perhaps be explained by the fact that the users were sometimes unsure of whether their original input had been correctly processed by the system.

Users often move from conversational to clear speech during repetition, something which partly explains some of the distinguishing features described in this section. The increase in average utterance duration is one of those features, and hyperarticulation is another. Inserted pauses were much more frequent in the repetitions than in the original utterances to the system, and they became increasingly frequent the longer the repetitive sequences lasted. Either the users of the system believed that they could resolve errors by means of modifying their articulation, and used phonetic features to indicate the contrast between original input and repetition, or they adapted their pronunciation without even being aware of it. For a given repetition, it is difficult to determine which explanation is the accurate one.

In the present study, some differences between the strategies used by adults and children were observed. Focus shifting in the repetitions occurred, but primarily among the adult users of the system. Adults also more often increased the duration of their utterance, spoke slower and inserted pauses between words. Children appeared to use other strategies for distinguishing a repetition from their original input to the system. As mentioned above, many of the children's repetitions were labeled as increased in loudness. This difference in adult and children strategies could have a number of explanations. It could be argued that while adults believe the system did not "understand" them the first time, children think the system did not "hear" what they were saying.

To examine some possible implications of the users' phonetic adaptations, all 452 utterances were analyzed in an experimental speech recognition test. In this experiment, a recognition lexicon was constructed by adding all missing words that occurred in these utterances to the lexicon that was used in the actual spoken dialogue system. The results from this preliminary experiment can be seen in Figure 8.4. The adult error rate was 37%, and the corresponding figure for children was 65%. Utterances with inserted pauses resulted in the highest figures for sentence error rate. These results indicate that for the experimental speech recognizer used in the August system (Ström 1997), computer-directed speech should be as neutral or unaffected as possible to be correctly recognized.

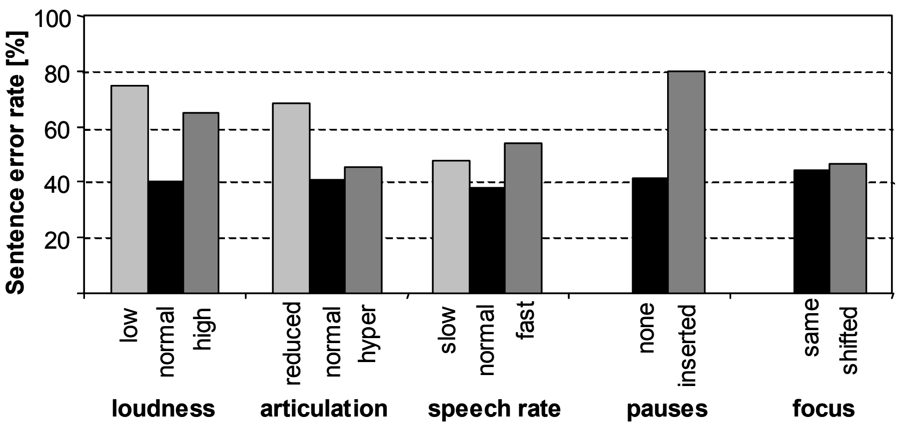


Figure 8.4 Sentence recognition error rates grouped by the linguistic features as labeled in the database.

The lowest recognition rates in the present test were observed in utterances with inserted pauses, as can be seen in the figure above. The explanation for this is that the recognizer used in this experiment had an insufficient model for silent segments within an utterance. Research has shown that the difficulty for speech recognizers trained on continuous speech to handle words spoken in isolation will not be solved by simply adding isolated speech to the training material (Alleva et al. 1997). The solution might be to have an isolated speech recognizer run in parallel, and let the dialogue manager predict which one to use depending on the context.

In this study, phonetic adaptations were examined by comparing original input to the August system with lexically identical repetitions of the same utterances. Characteristic features of hyperarticulate speech, such as increases in duration, decreases in speech rate and inserted pauses between words, were found in many of the repetitions. A tendency in the data was that children appeared to use partly different strategies than adults during error handling. In particular, children appeared to often increase the loudness of their speech when the dialogue system failed to understand them, something which adults did to a lesser extent.

8.2 Repetitive Sequences in the Pixie Corpus

Analyses of the August corpus indicated that adults and children use partly different strategies in their interaction with a spoken dialogue system. Because of the setup of this system, described in 7.1 above, these results cannot not be seen as conclusive. In the development of the Pixie system, we made a special effort to control the setup in order to get a well-defined speaker database. This made it possible to study longer sequences of child and adult dialogue behavior in this fully functional system. In the present study, phonetic adaptations in repetitive sequences taken from longer dialogues in the Pixie corpus are examined. Non-identical as well as identical repetitions are analyzed, and the differences between children and adults in terms of user strategies are discussed.

8.2.1 Background

Most spoken dialogue systems developed so far have been designed for adult users. However, a growing number of adolescents and children are likely to access speech based systems in the future. Studies have shown that children's speech is more variable in terms of acoustic-prosodic features as well as more disfluent when compared to adult speech (Lee, Potamianos and Narayanan 1999; Oviatt and Adams 2000). As a consequence, conventional speech recognizers, trained mainly on adult speech, have difficulty handling children's input (Wilpon and Jacobsen 1996; Potamianos, Narayanan and Lee 1997; Gustafson and Sjölander 2002). Moreover, it appears as if children employ partly different strategies when interacting with dialogue systems than adults do. Differences between children of different ages can also be seen. A Wizard-of-Oz study has indicated that younger children use less overt politeness markers and verbalize their frustration more than older children do (Arunachalam et al. 2001). It has also been shown that children's experience of using a system is improved if they can communicate with a system with a 'personality' and that they benefit from being able to choose from several input modalities (Narayanan and Potamianos 2002).

As speech technology develops, spoken dialogue systems will be used to perform increasingly difficult tasks. The demands on these systems to be robust enough to handle real-life environments and mobile users are also increasing. Most commercial applications developed to date have focused on users in a relatively quiet and controlled environment, such as a home or an office. During the last few years, however, spoken dialogue interfaces intended for public settings have also started to emerge. Examples are systems for information-retrieval tasks over the telephone and call routing (Aust et al. 1995; Gorin et al. 1997; Lamel et al. 1997; Lamel et al. 1998) and information kiosks in public places (Lamel et al. 1998; Gustafson and Bell 2000). In such systems, the outside environment is a factor which is potentially very difficult for the system to assess and handle.

When children begin to use speech-based systems, it is unlikely that they will want to sit in a quiet office and perform dictation tasks. Instead, kids will probably want to play portable games or access information in a variety of environments. The challenge involved in developing spoken dialogue systems for children is in a sense dual: children's spoken language is inherently more difficult to handle and the systems used by children are likely to be exposed to noisy real-life settings.

8.2.2 Data and Annotation

Details about the data collection for the Pixie system can be found in Chapter 7. From the Pixie corpus, 25,000 utterances were manually transcribed as well as labeled with tags for exaggerated pronunciation in terms of loudness and hyperarticulation. In the transcribed database as a whole the number of speakers is 2,885, and the average number of utterances per speaker is approximately nine.

To be able to examine whether the users' speaking rate increased or decreased during error handling, all utterances were acoustically analyzed. From the corpus of transcribed and labeled data, we took out 15,000 utterances from dialogues with more than five turns. The segmentation of the speech material into words and phonemes was achieved by means of an automatic alignment algorithm (Sjölander 2001; Sjölander 2003). The input to the auto aligner is a sound file and a verbatim transcription of the speech. The output consists of two tiers marking words in standard orthography, and phonemes, respectively. The phoneme tier is supplemented with lexical prosodic features such as primary and secondary stress and word accent type (i.e. accent I or II). The grapheme-to-phoneme conversion, as well as the lexical prosodic markup was accomplished with the KTH text-to-speech system (Carlson et al. 1982).

Our main interest in the current study was to examine users' error handling strategies in longer sequences of human–computer dialogues. To be able to examine phonetic features and dialogue strategies in greater detail, we made yet another selection. From the 15,000 aligned utterances, we randomly extracted 16 adult and 16 children speakers whose interactions with the system consisted of between 15 and 25 user utterances. The children in this sub-set of the corpus were between the ages of 9 and 12. Both the adult and child groups were gender balanced.

As previously mentioned, users engaged in two types of dialogues with Pixie. In the system-driven domain dialogues, speakers were asked to help Pixie perform certain tasks in the apartment. In the user-driven social dialogues, speakers could ask the agent questions about herself, the home of the future, or the exhibition. The corpus of 32 speakers was also manually labeled by the authors at the dialogue level. Each user utterance was assigned one of the following tags: *Normal*, *meta*, *error*, *repeat* and *rephrase*. *Normal* were all utterances that were part of the typical interaction with Pixie, *meta* were comments about the system or dialogue itself, *error* were utterances in which speakers refused to answer the system questions, *repeat* were verbatim repetitions of the previous utterance and *rephrase* all non-verbatim repetitions and rephrasings. Examples of labeled user utterances can be seen in Table 8.1-8.3 below.

We also manually labeled the dialogue for how the system's previous turn had affected user behavior, inserting the following tags for each user utterance: *Correct*, *rejected* and *wrong*. *Correct* was the label used when the user's previous turn was correctly handled by the system and *rejected* was used when the speech recognition confidence score was under a certain threshold which led to system prompts such as "I didn't hear/understand you." *Wrong* was used in the cases where there was a misrecognition which led to the wrong response.

Table 8.1 An error sequence (child)

Vart tog familjen vägen? *Where did the family go?*	normal
Vart tog familjen vägen? *Where did the family go?*	repeat
Vet du vart familjen tog vägen? *Do you know where the family went?*	rephrase
Vet du vart familjen tog vägen? *Do you know where the family went?*	repeat
Vet du vart familjen tog vägen? *Do you know where the family went?*	repeat
Vart tog familjen vägen? *Where did the family go?*	rephrase

Table 8.2 An error sequence (adult woman)

Jag vill ha vatten i badkaret *I want water in the bathtub*	normal
Sätt på kranarna *Turn on the faucets*	rephrase
Sätt på vattnet *Turn on the water*	rephrase
Sätt på sätt på vattenkranen *Turn on turn on the water faucet*	rephrase

Table 8.3 An error sequence (adult male)

Hur gammal är du? *How old are you?*	normal
Eh du har stora öron var kommer de ifrån? *Eh you have big ears where do they come from*	normal
Eh ja nu vet jag inte vad jag ska säga men *Eh well now I don't know what to say but*	meta
Vad gör du på fritiden? *What do you do in your spare time?*	normal
När stänger utställningen? *When does the exhibit close?*	normal
Vilka sorts saker äter du? *What sorts of things do you eat?*	normal

Finally, a subjective measurement of perceived speaking loudness was individually assigned to each utterance. In this case, an objective measure was not possible to use since the system was unsupervised and users stood at various distances from handheld microphones when interacting with Pixie. Instead, we listened to each user utterance and used the labels *low*, *normal*, *high*, *very high* and *scream*. Subjective measures of hypo- and hyperarticulation were also added, as well as labels for mispronunciation and 'silly voices'.

8.2.3 Results

Our analyses indicate that adults and children use partly different strategies during error handling with the Pixie system. When the dialogue fails, speakers often make one or several attempts to resolve the problem and make themselves understood. As can be seen in the dialogue example in Table 8.1 and in Figure 8.5, children often repeat the same utterance several times. Adults, on the other hand, tend to rephrase their original utterance instead of repeating it verbatim. In the user-driven social dialogues, this pattern is especially clear. When Pixie had failed to interpret their original utterance correctly, adults would attempt to rephrase it, as exemplified in Table 8.2. Alternatively, adults who could not make themselves understood would simply move on to the next query. The latter user strategy can be seen in Table 8.3 above.

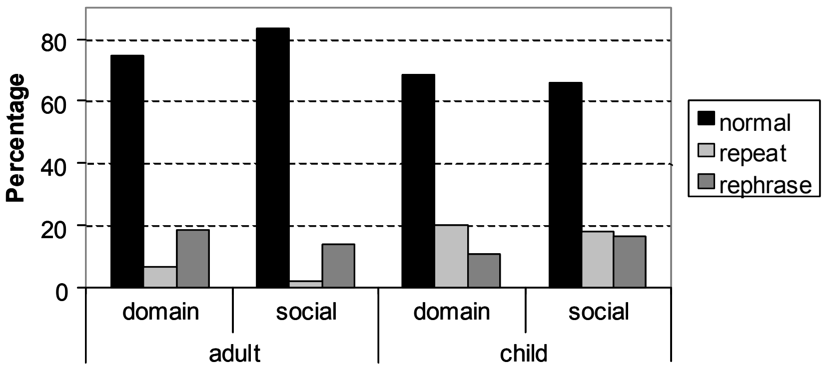


Figure 8.5 Percentage of all utterances in each category labeled as normal, repeat and rephrase

A closer examination of the utterances labeled as rephrase reveal some differences between adults and children within this group. When children rephrase their previous utterance they typically add or take away a non-content word. That is, they seldom or never modify the phrase structure or lexical content of the utterance. Table 8.1 contains an example of this type of repetitive sequence, where a child goes back and forth in her efforts to convey her message to the system. When adults rephrase a previous utterance, however, different patterns can be seen. Instead of being verbatim or near-repetitions, these sequences often contain greater changes in lexicon and/or phrase structure; see Table 8.2. In this sequence, the user tries several ways of expressing what she wants done, and modifies her lexical choices repeatedly.

During repetitive sequences, users often modify different acoustic-prosodic features of their speech. As shown in Table 8.4, children's utterances in all categories of repetitions were judged as hyperarticulated to a higher degree than adults'. The utterances labeled as *normal* were hyperarticulated almost twice as often for children when compared with adults. Half of the adults' verbatim repetitions were hyperarticulated, while the corresponding figure for children was 74%.

Table 8.4 Percentage of utterances labeled as hyperarticulated in different categories

	Normal	Repeat	Rephrase
Adults	16%	50%	28%
Children	29%	74%	41%

As can be seen in Figure 8.6, children more often use loudness as a way of distinguishing repetitions and rephrases from original utterances. About 70% of the children's verbatim repetitions were labeled as increased in loudness. While the adults' utterances were never labeled with *scream*, children rather frequently shouted during their interaction with the animated agent.

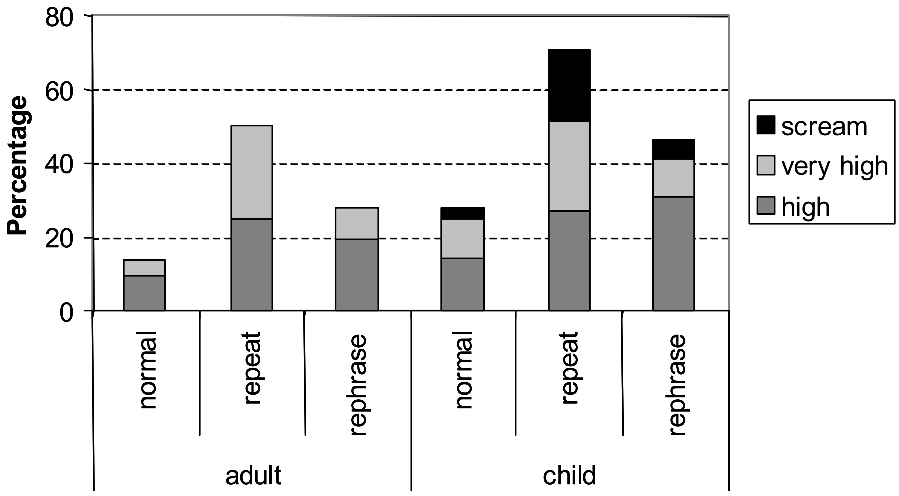


Figure 8.6 Perceived speaking loudness for different utterance types

In the corpus of 15,000 aligned utterances, verbatim repetitions were compared to original utterances. The automatic alignment algorithm had a high precision at the segmental level. However, the difficult acoustic environment with both visitors and other Pixie agents talking in the background made the silence detection quite poor. This meant that both original and repeated utterances had to be manually corrected. All in all, 392 repetitions from children and 387 repetitions from adults were analyzed. The repetitions were on average 29%

slower than the original utterances. For children, the repetitions were slightly more than 30% decreased in speaking rate overall. For adults, the repetitions in the domain dialogues were also about 30% slower, while the repetitions in the social dialogues were 22% slower.

The system's previous utterance also affected user strategies in the Pixie corpus. When the system had said "I didn't understand/hear you" in the previous turn, the speaking rate was decreased for both adults and children. Again, this adaptation of speaking rate was more exaggerated for children. This decrease in speaking rate is related to the fact that the repeated utterances often were hyperarticulated. When longer sequences were analyzed, original utterances appearing at the beginning and end of the dialogues were not significantly different in terms of speaking rate. The adaptation of speaking rate that occurs is local, primarily affecting the utterance immediately following a problematic turn.

Children sometimes found it difficult to pronounce some of the options given to them in the system-driven domain dialogues. 11% of the children's utterances in certain difficult dialogue turns contained mispronunciations. Finally, children speakers occasionally modified their manner of speaking by using 'silly voices'. Extreme modifications of voice quality and cartoon-like speech imitations characterize these utterances, which occurred in 4% of all cases.

8.2.4 Discussion and Summary

Collecting spoken dialogue data in a public environment is a challenging task. It is difficult to maintain control of all variables as large quantities of human–computer dialogues are recorded in a stand-alone system. In the Pixie system, the smart cards used for registration and interaction proved to be a robust solution to the problem of assigning user identity, age and gender. In a noisy environment with multiple dialogue systems running in parallel, speech recognition and silence detection become problematic. In order to develop a speech recognizer that is robust enough to be used in public, it is necessary to train new acoustic models based on data from real-life settings. However, speech detection and automatic alignment methods are less reliable for this kind of data. Until better models are available, we must use labor-intensive manual methods for preparing speech data from public settings for training purposes.

Children and adults react to system errors in different ways. In the repetitive sequences, this can partly be explained by the fact that it appears to be easier for adults to come up with various ways of rephrasing their utterances. Children have not yet perfected their language skills, and sometimes fail to come up with an alternative way of expressing a request. It is probably more difficult for children to modify lexical content and syntactic structure, and they tend to repeat the same thing over and over again or make only minor modifications to their previous utterance. For both adults and children, verbatim repetitions are often hyperarticulated, increased in loudness and longer in duration. However, children's pronunciation is more exaggerated in these respects. Furthermore, when the system fails to understand them adults often move on to the next question in the social dialogues. Children are often more persistent in trying to get the system to understand their questions in both domain and social dialogues.

Several aspects of the Pixie system contributed to the patterns of dialogue behavior described above. The acoustic environment is clearly one such factor, since the general noise level in the exhibition area was quite high. Children visitors often came in larger groups, with many kids simultaneously talking to Pixie and other visitors in the room. This can partly explain why children hyperarticulate and raise their voices to such a degree during their interaction with the animated agent. Furthermore, the dialogue design of the system was not tailored for young children and it was sometimes hard for these users to know what to say or (in the system-driven part of the dialogue) to pronounce the options available. Another factor that may have contributed to the results presented is that the users that were picked out were taken from a group who engaged in longer interactions with Pixie. It is possible that neither adults nor children in this group were representative of the entire population of Pixie users. In particular, the younger users who persevered in trying to communicate with the system despite recurrent errors may belong to a group of persistent children, who were perhaps more immersed in the dialogue.

In this study, we have described how speakers adapt their language during error resolution when interacting with the animated agent Pixie. Investigations comparing adults with children between the ages 9 and 12 indicate that the two groups of subjects use partly different strategies during error handling with a publicly exhibited spoken dialogue system. Children often repeat the same utterance verbatim several times, while adapting their speech phonetically and prosodically. 74% of the children's repetitions were labeled as hyperarticulated, and 70% were increased in loudness. Apart from phonetic adaptations, adults also have access to other dialogue strategies, and often modify their lexicon and/or syntax to meet what they believe to be the limitations of the system. More research is needed to increase our knowledge of the differences between adult and children behavior during error resolution in spoken dialogue systems. It is worth considering that children behave differently than adults do, and that we should develop spoken dialogue systems that better respond to their needs and interests.

8.3 Speaking Rate in the Cloddy Hans Corpus

As discussed in Chapter 4 above, human dialogue participants often adapt their manner of speaking to that of the other participants. This adaptation takes place at several linguistic levels, allowing phonetic, prosodic, lexical, syntactic and pragmatic aspects to come into play. For example, if one person speaks in a low voice, the other participants might start whispering as well. In this chapter, we address the question of whether people who interact with a spoken dialogue system are likely to adapt their speech at the prosodic level.

In the preceding sections of this chapter, we examined how speakers adapted their manner of speaking during repetitive sequences in the August and Pixie corpora. In the present study, we investigate whether users who interact with the animated agent Cloddy Hans adapt their speech rate to match that of his spoken output. Subjects were divided into two groups, and were assigned to interact with either a 'fast' or 'slow' version of the system. The users' input utterances were analyzed to see if their speech rate had been affected.

8.3.1 Background

When people engage in conversation with spoken dialogue systems, it is sometimes necessary to influence their manner of speaking in order to facilitate their interaction. State-of-the-art speech recognizers are predominantly trained on normal speech and have difficulty handling either exceedingly slow and hyperarticulated, or fast and reduced speech. Explicitly instructing users on how to speak, however, can make the human–computer interaction stilted and unnatural. If it is possible to affect users' speaking rate while maintaining the naturalness of the dialogue, this could prove useful in the development of future human–computer interfaces. Users could thus be subtly influenced to adapt their speech to better match the current capabilities of the system, so that errors can be reduced and the overall quality of the human–computer interaction is improved. At the same time, speakers are allowed to express themselves freely and naturally.

Chapter 4 addressed the question of how people interacting with spoken dialogue systems adapt their manner of speaking in ways that are appropriate in human–human interaction. However, we observed that this behavior is not always suitable for human–computer interaction. A typical example is what users of dialogue systems do when they have been misunderstood by the system. Although users are not interacting with another human being, they often hyperarticulate, lower the speech rate, increase the loudness, insert pauses between words, et cetera (Oviatt et al. 1998a). That is, they use the same means to increase intelligibility as they would in dialogue with other humans. For the speech-understanding module of a spoken dialogue system, unfortunately, this strategy appears to have the opposite effect. As discussed in Chapter 5, hyperarticulated speech has been shown to elevate speech recognition failures in human–computer interaction (Rhyne and Wolf 1993; Yankelovitch, Levow and Marx 1995). Efforts to model users' hyperarticulate speech during error resolution may result in the development of future systems with an improved ability to handle such input (Oviatt et al. 1998c; Soltau and Waibel 2002). However, speech recognizers of today are almost entirely trained on neutral, unaffected speech and are ill equipped to interpret speech which is hyperarticulated, emotionally colored or excessively fast or slow. Making users aware of system limitations and telling them how to speak can make the human–computer interaction seem less natural. Moreover, attempts to instruct users to 'speak naturally' to make their language correspond to that of the speech recognizer's training model have not been successful (Shriberg et al. 1992).

Previous studies with simulated systems have shown that children adapt both their response latencies (Darves and Oviatt 2002) as well as their amplitude (Coulston, Oviatt and Darves 2002) to that of their conversational partner, in this case different TTS voices.

8.3.2 Data and Annotation

Details regarding the data collection with the NICE system can be found in Chapter 7. In brief, 16 subjects interacted with the system's full-size fairy tale character Cloddy Hans, who asked the users for help in solving a task. The users had been given a pictorial scenario which involved asking Cloddy Hans to move a series of geometrical objects, and were informed that they should communicate using spontaneous spoken language. As mentioned in 7.4, subjects either interacted with a 'slow' or 'fast' version of Cloddy Hans. The speaking rate of the feedback prompts in the 'slow' version had been decreased by 30%, and the speaking rate of the prompts used in the 'fast' version had been increased by 30%.

The total number of recorded user utterances was 297. Each dialogue consisted of 6 tasks (i.e. there were 6 colored geometrical figures to be moved) and 16 to 27 user turns. The total number of word tokens in the corpus was 2,173. The entire corpus was orthographically transcribed, and all user turns were tagged with information on position in the dialogue, type of user turn, previous system output, etc. At the word level, the user utterances were also labeled for lexical content, distinguishing between 'color', 'shape' and 'other' words. However, for the purpose of this particular study, only a subset of the user turns was of interest, namely the first turns in each task, and repetitions and rephrases of these.

A total of 130 user turns consisting of acknowledgements such as "ok" or "yes" and a few erroneous utterances were excluded from the subsequent analyses. The average length of the remaining 167 utterances was about 6 seconds.

Table 8.5 below shows an example dialogue from the corpus with labeling of position and type of user turns. For example, the utterance labeled '2.1_original' is the first turn of the second task in the dialogue, '4.2_repeat' is a (near) verbatim repetition of the first turn of the fourth task, and '4.3_rephrase' is a turn where the user repeats his previous turn but in different words. The turns labeled 'ack' are acknowledgements of the system's previous explicit question, which were excluded from the speech rate analyses.

Table 8.5 Excerpt from a dialogue between a user and Cloddy Hans.

User	**User turn**	**System**
Eh röd tr eh röd fyrkant ta röd fyrkant eh vänster *Eh red tr eh red square take red square eh left*	1.1_original	Det där förstod jag inte *Now I didn't understand*
Ta röd fyrkant och placera ovanpå blå triangel *Take red square and place on top of blue triangle*	rephrase	Vill du att jag ska flytta den röda fyrkanten till den blåa trekanten? *Do you want me to move the red square to the blue trigon?*
Ja *Yes*	Ack	Ok *Ok*
Flytta den blåa cirkeln till den gröna fyrkanten *Move the blue circle to the green square*	2.1_original	Det där förstod jag inte *Now I didn't understand*
Blå cirkel till grön fyrkant *Blue circle to green square*	2.2_rephrase	Vill du att jag ska flytta den blåa cirkeln till den gröna fyrkanten? *Do you want me to move the blue circle to the green square?*
Ja *Yes*	Ack	Ok *Ok*
Grön triangel till röd cirkel *Green triangle to red circle*	3.1_original	Vill du att jag ska flytta den gröna trekanten till den röda cirkeln? *Do you want me to move the green trigon to the red circle?*
Ja *Yes*	Ack	Ok *Ok*
Den gula cirkeln till den gröna trekanten *The yellow circle to the green trigon*	4.1_original	Vill du att jag ska flytta den blåa fyrkanten till den röda cirkeln? *Do you want me to move the blue square to the red circle?*
Nej den gula cirkeln till den gröna fyrkanten *No the yellow circle to the green trigon*	4.2_repeat	Vill du att jag ska flytta den blåa cirkeln till den gröna trekanten? *Do you want me to move the blue circle to the green trigon?*
Nej ljusgrön cirkel till grön trekant *No lightgreen circle to green trigon*	4.3_rephrase	Det där förstod jag inte *Now I didn't understand*

The acoustic analysis involved a quantitative estimate of user speaking rate calculated by taking the average normalized segment duration across each word. A standard z-score technique (Wightman et al. 1992) was used to normalize for inherent duration, and the means and standard deviations used in the normalization were taken from a database of all segments from all speakers in the experiment. However, a prerequisite for this estimate was a segmentation of the speech material into words and phonemes. This was achieved by means of an automatic alignment algorithm (Sjölander 2001; Sjölander 2003). The input to this aligner is a speech file and a verbatim transcription of the speech. The output consists of two tiers marking words in standard orthography, and phonemes, respectively. The phoneme tier is supplemented with lexical prosodic features such as primary and secondary stress and word accent type (i.e. accent I or II), thus enabling separation of stressed and unstressed vowels etc., in the z-score normalization.

8.3.3 Results

Figure 8.7 illustrates some of the speaking rate variations in the user utterances. As can be seen, the two groups of subjects started out with similar speech rates (cf. 1.1_original in Table 8.5). At this point in the dialogue, they had not yet heard Cloddy Hans's voice and could not have been affected by his speech. As illustrated in the dialogue example above, the subjects were uncertain on how to express themselves here and many of the utterances coded as 1.1_original were low in speech rate, fragmented and contained disfluencies. Subsequently, subjects in the 'fast' and 'slow' groups began to diverge in terms of speaking rate. The users' second utterance (2.1_original) was deliberately rejected by the Cloddy Hans, who said "Now I don't understand" and shrugged his shoulders in resignation. Subjects reacted to this by either repeating their original utterance verbatim or by rephrasing it. An interesting observation was that the repetitions in this dialogue context (2.2_repeat) were found to be increased in speaking rate while the rephrases (2.2_rephrase) were decreased in speaking rate. In contrast, later on in the dialogue the system deliberately misunderstood the user's utterance completely. Cloddy Hans then responded with a feedback prompt which revealed that he had failed to understand either of the user's references to geometrical objects. In this case, both repetition (4.2_repeat) and rephrase (4.2_rephrase) were pronounced slower.

A statistical test (ANOVA) was used to examine the effects of system speaking rate, user turn, and lexical content. The dependent variable was the estimate of user speaking rate described above, and the independent variables were:

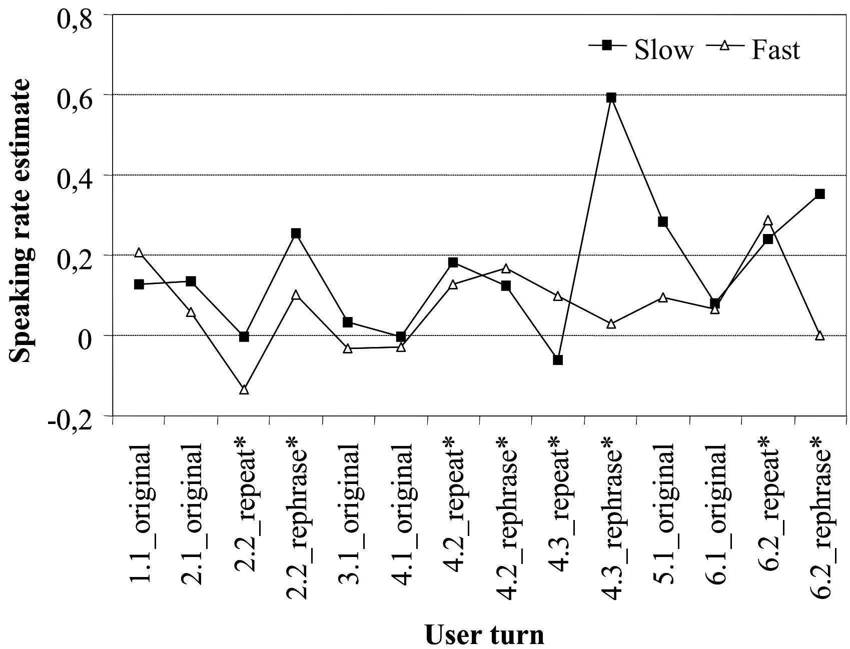


*Figure 8.7 Effects of system speaking rate (fast vs. slow) on user speaking rate (measured as the average z-score normalized segment duration across words in std. devs.) in different user turns. Users either rephrased or repeated turns but not both, these turns are marked with *.*

- system speaking rate (fast vs. slow)
- user turn
- lexical content (color vs. shape vs. other)

The test showed a significant effect of system speaking rate in the expected direction (fast<slow)[4]. In addition, there were significant effects of user turn and of the interaction between speaking rate and user turn. That is, turns differed significantly in speaking rate, and the effect of system speaking rate differed between turns. Finally, there was a significant effect of lexical content.

Another statistical test (Bonferroni pairwise comparison) on the effect of lexical content, showed that 'color' words were pronounced significantly slower than 'shape' and 'other' words, while there was no significant difference between 'shape' and 'other' words. None of the other effects were significant.

[4] The figures for the statistical computations can be found in Bell, Gustafson and Heldner (2003).

Furthermore, another ANOVA was used to examine the effects of silent pauses within user utterances. The dependent variable here was the absolute duration of the silent pauses, and the independent variables were:

- system speaking rate
- user turn

There was a significant effect of user turn, but neither speaking rate, nor the interaction of speaking rate and user turn was significant. However, because the number of occurrences was small, the power to detect statistical significances was fairly low.

8.3.4 Discussion and Summary

Our results support the hypothesis that users interacting with the 'slow' version of the system speak slower than those who interact with the 'fast' one. The results also reveal substantial variations in speaking rate that can be attributed to the dialogue context. A tendency in our data is that when the dialogue is successful, an increase in speech rate is elicited. Local effects on speech rate, such as a slow user utterance after a system misunderstanding, appear to be transient and quickly passing. Once the system seems to understand their input again, users return to their normal manner of speaking. Among the local convergence effects, we found that repeats and rephrases tended to be slower after a system turn that echoed the referents of the user's previous input in a completely erroneous way. In these cases, users probably spoke slower as a result of an increase in cognitive load caused by the simulated error. Moreover, our analyses showed that the lowering of speech rate mainly affected the 'color' words in the dialogues. Users often modified the 'shape' words lexically by exchanging a word for a near-synonym. In the course of the dialogue, there was a lexical convergence effect where users often conformed to Cloddy Hans's vocabulary. It was more difficult for the users to come up with alternatives to the 'color' words, and only two such instances were found in the corpus. Furthermore, color words constituted contrastive material in the sense that all misunderstood turns concerned color and shape or color alone, and never shape alone. Finally, the analyses revealed that the lowered speech rate did not affect the within-utterance silent pauses significantly.

One possible interpretation of the results is that users modify their speech according to their current model of the system's input understanding capabilities. Two factors seem to influence the users' model: the system's output speech rate and the system's ability to handle the previous utterance(s). While successful turns elicit an increase in speech rate for both groups of users, subjects who interact with the 'slow' version of the system are affected to a lesser extent. On the other hand, while both groups of users react to simulated system errors by speaking slower, users interacting with the 'fast' Cloddy Hans tend to do this to a lower degree.

Post-experimental discussions confirmed that subjects were aware of the fact that they were adapting their language at the lexical level. Several subjects spontaneously mentioned the fact that they deliberately modified their vocabulary to match that of the system. However, adaptation of speaking rate was not mentioned as a strategy consciously used by the subjects. The tendency in our data that subjects decreased their speaking rate can partly be attributed to the general impression of the animated agent. The fact that Cloddy Hans's speaking style and general appearance implied a certain limitation in his intellectual abilities was probably relevant.

In this study, we wanted to investigate whether users could be influenced to adapt their speech rate when interacting with an animated character in a simulated spoken dialogue system. The experiment confirmed that the subjects adapted to the speaking rate of the system, although they varied their speaking rate substantially in the course of the dialogue. In particular, problematic sequences where subjects had to repeat or rephrase the same utterance several times elicited slower speech.

Chapter 9

Lexical and Syntactical Adaptations

People who engage in dialogue have been observed to coordinate their language with their interlocutors at the lexical and syntactical levels, as discussed in Chapter 4. In human–computer interaction, there are similar patterns where the system's output influences user behavior. In this chapter, we describe empirical studies which investigates how users adapt their language lexically and syntactically during spontaneous interaction with a spoken dialogue system. The work presented here is based on studies of two corpora.

The first part of this chapter is based on studies of the August database. The major parts of this work, carried out in collaboration with Joakim Gustafson, have previously been published in the *Proceedings of Eurospeech* (Bell and Gustafson 1999b) and in an article in *Natural Language Engineering* (Gustafson and Bell 2000). As previously mentioned, Joakim Gustafson made a great contribution to the development and implementation of the August system. Both authors contributed equally to all aspects of tagging and labelling the corpus, as well as analyzing the results.

In the second study, which has not previously been published, occurrences of lexical adaptation in the Cloddy Hans database are analyzed. Data collection was carried out by Linda Bell in collaboration with Joakim Gustafson and Mattias Heldner, as described in Chapter 8. Labelling and analysis of data for the present study was done by Linda Bell and Joakim Gustafson in collaboration.

9.1 Lexical and Syntactical Adaptations in the August Corpus

As noted in 3.1 above, early studies with text-based dialogue systems suggested that computer-directed language would be simple and command-like in style, and that it can be characterized by a restricted vocabulary and the absence of pronouns (Guindon et al. 1987; Kennedy et al. 1988; Dahlbäck 1991). In this study, some of the linguistic strategies employed by the users of the August spoken dialogue system were investigated. The aim was to examine whether the type of user adaptation that is characterized by short utterances, avoidance of anaphora and a limited vocabulary is prevalent in the current database. Lexical aspects of the August corpus are discussed and word and utterance statistics are presented. Adaptations in syntactical structure are also investigated, especially in the context of repetitive utterances spoken during error resolution.

9.1.1 Background

When people interact with spoken dialogue systems, they try different approaches to make their communication successful. For instance, people who interact with computers often adapt their linguistic behavior to meet the demands of the system. As discussed in Chapter 5, users are even more likely to modify their language when the interaction with the system does not run smoothly, i.e. during error resolution.

Most experienced users of spoken dialogue systems are aware of system limitations in terms of lexicon, and are likely to modify their vocabulary to match what they believe to be the system's capabilities. Inexperienced users, on the other hand, may expect spoken dialogue systems to handle large or even unlimited vocabularies. It has also been shown that people who interact with computers tend to reuse words and structures used by the system itself (Brennan 1996; Gustafson et al. 1997). It is assumed that if a specific lexeme, phrase or idiomatic expression occurs as system output, it should be possible to recognize as user input as well. Other user strategies include keeping utterances short and avoiding anaphoric usage of pronouns. When the interaction with the system goes wrong, users are likely to attempt more advanced strategies for resolving errors.

The system interface design also contributes to people's expectations and may indirectly influence their linguistic behavior as they interact with the system. A central feature of the present system was that August's synthetic face was intended to appear human-like and that he was given a 'personality.' The animated agent's synthetic speech output was lip-synchronized and extra attention had been given to make the prosody of the voice sound natural. Moreover, August's face was given a varied set of extra-linguistic gestures and expressions, such as eyebrow-movements, movements of the head and movements of the eyes and eyelids indicating that he was "thinking" (Lundeberg and Beskow 1999). The animated agent's human-like appearance made us anticipate that the users would want to exchange greetings and socialize with him. The ability to handle and respond to some of these social utterances was built into the system. It could be argued that the animated agent's face, with the above-mentioned features, influenced the way in which the users interacted with the August system.

9.1.2 Data and Annotation

The August corpus, described in Chapter 7 above, consists of 10,058 utterances of computer-directed speech. As previously mentioned, all utterances were transcribed orthographically and some basic speaker characteristics were manually labeled, so that men, women and children among the users of the system could be roughly divided into groups. The total number of users was estimated to 2,685, out of which about 50% were men, 26% women and 24% children.

All utterances and words were sorted according to their frequency in the database. The entire corpus was part-of-speech tagged and parsed using automatic tools (Karlsson 1992; Carlberger 1999). All unique words were extracted from the database and the number of words per utterance was computed. To be able to compare the August corpus with another Swedish database of computer-directed speech, we also studied the Waxholm corpus. A Wizard-of-Oz version of the Waxholm system, which provided information on boat traffic, had been used to collect a database which had been linguistically analyzed (Bertenstam et al. 1995b). As a baseline for investigating aspects such as general complexity and word coverage, we also considered the KTH corpus. This is a labeled database consisting mainly of Swedish newspaper text, approximately 150 million words (Bertenstam et al. 1995b).

Those utterances including words that had been given the part-of-speech tag *indefinite pronoun* were excerpted from the corpus. A number of these indefinite pronouns were believed to refer to back to previous turns in the dialogue and could constitute instances of anaphora. Because these expressions can be ambiguous in Swedish, they had to be manually labeled to make sure only genuinely anaphoric references were included. The authors listened to the utterances which contained indefinite pronouns, as well as the utterances spoken just before and after these.

All utterances in the corpus were divided into the two categories *information-seeking* and *socializing*, respectively. Typical utterances in the information-seeking category include "Hur mycket är klockan?" ('What time is it?') and "Var finns det restauranger i Stockholm?" ('Where can I find restaurants in Stockholm?')

while the socializing category can be exemplified by utterances such as "Hej August!" ('Hello August!') and "Hur gammal är du?" ('How old are you?'). This categorization of the corpus is described in greater detail in Section 10.1 below.

As discussed in Chapter 8 above, users of a spoken dialogue system who fail to make themselves understood often repeat or rephrase what they have just said. Sequences of near-repetitions were therefore believed to be interesting from the point of view of lexical and syntactic user adaptation. Exact and approximate repetitions together constituted 12% of all utterances in the current database. Results from a phonetic investigation of lexically identical repetitions in the August database is reported in Section 7.1 of this thesis. However, only non-identical repetitions displaying lexical variation were examined in the present study. This sub-set made up 4% of all utterances in the corpus, or 402 utterances. It appeared likely that any patterns of lexical and syntactical adaptation would emerge as users rephrased their previously misunderstood utterance to the August system. Consequently, the purpose of studying non-identical repetitive sequences in this context was to get a clearer picture of any changes of lexical and syntactical patterns that occurred during error resolution. In the categorization of non-identical repetitions, the following features were used:

- The exchange of one lexical item for another
- Change of word order
- Insertion/deletion of a word or phrase
- Increase/decrease in syntactic complexity

In addition, a small number of repetitions which could not be categorized according to the features above were labeled as *other*.

9.1.3 Results

The total number of words in the database was 39,230, out of which 23,604 words belonged in the information-seeking category and 15,626 in the socializing category. 2,918 word forms in the corpus were unique and half of them were hapaxes, i.e. words that occurred only once in the corpus. In the Waxholm spoken dialogue system, the number of unique words was 600 (Bertenstam et al. 1995a). The 200 most frequently used words in the Waxholm database covered 92% of all words. By contrast, the 200 most frequently occurring words in the August database covered 81%. This can be seen in Figure 9.1 below. These differences can partly be explained by the fact that the domain of the Waxholm system is closed when compared to that of the August system.

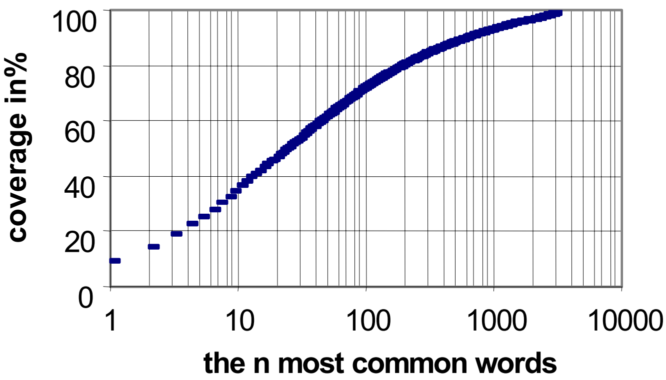


Figure 9.1 Word coverage in the August corpus as a function of frequency.

Some differences between the lexicons of the two utterance categories in the corpus can be seen. 1,431 of the words occurred only in the information-seeking category, and 20% of these were hapaxes. The corresponding figure for the socializing category was 852 words, out of which as many as 67% were hapaxes. The overlap between the two categories was 632 words, many of them common function words and main and auxiliary verbs. The most frequently occurring words in the August database were then compared with the most frequent words in the KTH corpus of newspaper text. When the 200 most frequent words in the KTH corpus had been listed, it could be noted that 14 of those words were not present in the August corpus. The corresponding figure for the Waxholm system was 75 words. However, 99 of the words that occurred in the August corpus could not be found in the KTH corpus of 150 million words. Most of these words were names, expletive expressions (often including swear words) and nonsense words.

The average number of words per utterance was 3.8 for men as well as children and 4.3 for women. The average utterance in the database was thus rather short. An interesting finding was that the utterances did *not* become shorter and more "telegraphic" as the interaction with the system went on, but rather retained the same length or even became longer. As illustrated in Figure 9.2, there were no relevant differences between the user groups in this respect.

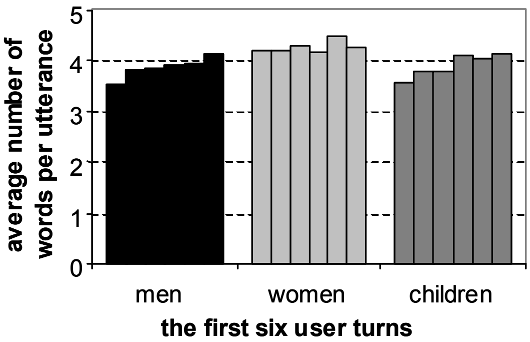


Figure 9.2 The average number of words in the first six user turns.

As indicated by the average number of words per utterance, there were not many instances of syntactically complex structures in the August corpus. The parsing of the corpus resulted in lists of phrase level structures, for example (NP) (VP) (NP) (PP) (Carlberger 1999). It turned out that more than half of all the utterances in the database could be covered by ten such syntactical structures. The syntactical patterns of the phrase structures of the utterances found in the socializing category were more uniform than the ones in the information-seeking category. 780 structures occurred only in the information-seeking category while the corresponding figure for the socializing category was 374. Most of these, however, occurred only once.

The overlap between the two categories, 188 phrase structures, covered 80% of all the utterances in the database. Most utterances contained only one verb phrase. Furthermore, in the cases were two verb phrases were found, the heads of these phrases were usually one main and one auxiliary verb. A majority of utterances were made up of one single clause and coordinated or subordinated clauses rarely occurred. Most utterances in the database can be described as syntactically simple.

Indefinite pronouns that were used anaphorically occurred in 1.6 % of the utterances in the corpus. Even though these anaphorically used pronouns were not very frequent, it is interesting that they occur. In the August corpus, they are used to talk about the system itself and refer back to previous turns in the dialogue. In 10% of the meta-utterances, anaphoric indefinite pronouns could be found. An example of a dialogue with an anaphoric pronoun can be seen in Table 9.1 below. This excerpt from the corpus illustrates the anaphoric use of pronouns that refer to entire utterances rather than to previously mentioned noun phrases. The anaphoric pronouns in the August database often had an entire utterance as their antecedent. Moreover, the dialogue excerpt in Table 9.1 below indicates the difficulty of handling utterances that refer back to previous turns in the discourse or the dialogue itself.

Table 9.1 An excerpt of a dialogue from the August corpus with an anaphorically used pronoun in the last user turn. The dialogue in its entirety went on for 15 turns.

User	August
Vad heter du? *What is your name?*	Jag kallar mig Strindberg, men egentligen har jag inget efternamn *I call myself Strindberg, but I don't really have a surname*
När föddes du? *When were you born?*	Strindberg föddes år 1849 *Strindberg was born in 1849*
Vad jobbar du med? *What do you do for a living?*	Jag kan svara på frågor om Strindberg, KTH och Stockholm *I can answer questions about Strindberg, KTH and Stockholm*
Hur många människor bor i Stockholm? *How many people live in Stockholm?*	I tätorten Stockholm bor det över en miljon människor *More than a million people live in the Stockholm area*
Trivs du här? *Do you like it here?*	Man ska inte kasta sten i glashus *People who live in glass houses should not throw stones*
Ja, **det** var smart sagt *Yes, **that** was a smart thing to say*	Hej, jag kommer från inst. för tal, musik och hörsel *Hello, I come from the dept. of speech, music and hearing*

Sequences of near-repetitions (e.g. Hi, what's your name? / Hello, what's your name?) were analyzed to see how users adapted their syntactical and lexical patterns during error resolution. As mentioned above, this sub-group made up 4% of the utterances in the entire database. Table 9.2 is an overview of the changes in the near-repetitions when compared to the original input to the system. As can be seen, the most commonly observed modification from original to repetition is the exchange of one lexical item for another. A less frequently used term was often exchanged for a more common one, as in the sequence "Do you have a *kitty*?"/ "Do you have a *cat*?" Generally, the users seem to be testing different strategies in their attempts to interact successfully. When the subjects repeated the same utterance more than once, they often alternated one feature such as increased/decreased syntactic complexity or insertion/deletion of a word or phrase (e.g. "How are you?" / "How are you, August?"/ "How are you?"). This pattern of linguistically contrastive pairs occurred in 41% of these sequences.

Table 9.2 Changes in % of all lexically different repetitions

	Changed lexical item	Changed word order	Inserted word/phrase	Deleted word/phrase	More complex	Less Complex
Men	29,2	11,0	15,8	11,5	9,6	13,9
Women	41,0	13,0	10,0	16,0	4,0	9,0
Children	34,5	12,6	12,6	13,8	12,6	6,9

9.1.4 Discussion and Summary

It has been claimed that people who interact with computers will adapt their language at the lexical and syntactic levels. For instance, computer-directed language has been said to be characterized by features such as a command-like syntax, an avoidance of pronouns and a restricted vocabulary.

The following tendencies were observed in the August database: First, considering that the users had not been instructed about how to interact with the system and that the system had more than a single domain, the number of unique words in the corpus is not very large. This indicates that the speakers used many common words and expressions. However, a number of odd words and expressions were found in the database. These words occurred more often in the information-seeking category than in the socializing category. Second, a relatively small number of phrasal combinations covered almost all of the syntactic structures in the database. The utterances were mostly quite short and simple and with few exceptions did the users seem to avoid syntactically complex structures. However, it is noteworthy that the users' utterances did not become increasingly simple or 'telegraphic' in the course of their dialogue with the system. Most of the complex structures in the corpus occurred as people referred back to the discourse during error resolution or when they were testing the limits of the system. Third, anaphorically used pronouns did not occur frequently in the database. When these anaphoric constructions were found, nonetheless, their antecedents were entire utterances rather than single noun phrases. An interesting observation was that people not only referred to previous turns in the dialogue but also talked about the system itself. Finally, user strategies in non-identical repetitive sequences included lexical as well as syntactical modifications. These features were often used in a contrastive manner.

In this study, we focused on lexical and syntactical aspects of the August database. In general, the language in the corpus was found not to be very complex at the syntactic level. Although utterances were mostly short, subjects did not use command-like language. Anaphorically used pronouns occurred in the corpus, albeit infrequently. Results from studies of non-identical repetitive sequences showed that people modified different aspects of their utterances during repetition, often by simplifying lexical content or syntax.

9.2 Lexical Adaptation in the Cloddy Hans Corpus

In this section, the question of how users adapted their vocabulary to match that of the system output in the Cloddy Hans corpus is addressed. In the analysis of the database, it became obvious that the subjects' lexical choices were more varied than we had anticipated. However, it was also apparent that the users were strongly influenced by the system's way of referring to objects and its choice of verbal constructions. The speakers' lexical choices at the beginning and end of their dialogues with Cloddy Hans are compared and discussed.

9.2.1 Background

As discussed in Chapter 3 above, humans coordinate their lexical choices in dialogue with those of their interlocutors. A convincing way of explaining this *lexical entrainment* effect has been put forward by theorists in the framework of grounding theory (Clark 1996; Brennan, 1996). Brennan and Clark (1996) note that the point of this collaborative process is not only being as efficient and terse as possible. In human–human dialogue, participants establish conceptual pacts which they stick to in later references even if this sometimes means being overinformative. For instance, if two speakers discuss objects with different geometrical shapes and one of them refers to a specific object by calling it *the round blue ball*, the listener may accept and even adopt this way of referring to it in subsequent dialogue turns. Branigan et al (2000) argue that there are functional benefits which explain why dialogue participants coordinate their language at the lexical and semantic levels. These benefits include a reduced computational load for the speaker, and a decreased risk for misunderstanding in the dialogue.

Lexical adaptation in the context of spoken dialogue systems is discussed in Chapter 4 above. The equivalent of lexical entrainment in human–computer dialogue, called *lexical convergence*, is a unidirectional process in which the human dialogue participant adopts the vocabulary of the system she is interacting with (Brennan 1996). Brennan argues that users are aware of the fact that spoken dialogue systems are not able to discuss what terminology to use, and tend to yield to the system's suggestions. An additional explanation might be that speakers are accustomed to adapting their language to meet the needs of different 'at risk' listeners, and modify their lexical choices without even thinking about it.

9.2.2 Data and Annotation

The Cloddy Hans corpus, described in Chapters 7 and 8 above, contains dialogues between 16 volunteer subjects and a Wizard-of-Oz version of a spoken dialogue system. Subjects interacted with the system's full-size fairy tale character Cloddy Hans, who asked the users for help in solving a task. The users had been given a pictorial scenario which involved moving geometrical objects in a particular order, and were informed that they should do this by interacting with the agent using spontaneous spoken language. The pictorial scenario can be seen in Figure 9.3. To help the users along, a prerecorded verbal presentation of the task was also played out loud to them. In this presentation, users were told they could help the agent by taking the geometrical objects and 'placing' them ("föremålen ska placeras" in Swedish) on shelves. The passive construction of the verb *placera* ('to place'), which sounds slightly awkward in this context, was deliberately used to make it more difficult for the subjects to reuse this terminology. Once the actual interaction with the system commenced, the system's output consisted of prerecorded prompts. All the way through the dialogues, the system gave users feedback using the verb *flytta* ('move'). As described in Chapter 8, users interacted with a slow or fast version of Cloddy Hans. However, in both of his appearances, the animated agent was clumsy and rather slow to understand.

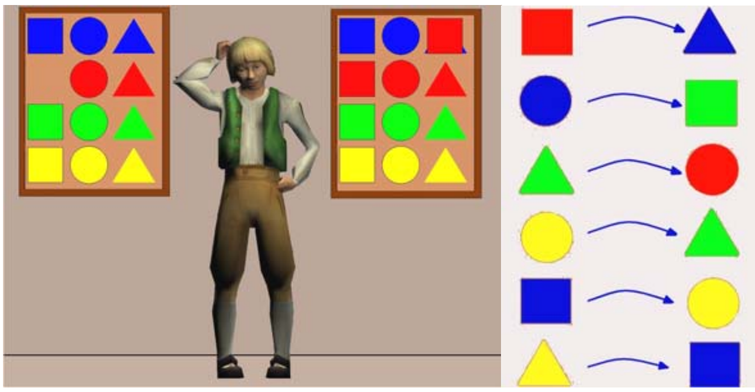


Figure 9.3 The graphical user interface and pictorial scenario.

The use of a pictorial scenario with geometrical shapes made it possible to avoid priming users in their choice of referential expressions to these objects. Instead, their initial way of describing these shapes was based on their immediate impression of how they looked. When the system later gave feedback on which objects to move where, the lexical items *trekant* ('trigon') and *fyrkant* ('square') were consistently used. These terms were chosen because they were believed to be neutral and frequently used. Table 9.3 shows a typical exchange between Cloddy Hans and a user:

Table 9.3 The first turns of a dialogue between a user and Cloddy Hans. Examples of lexical adaptations are marked in bold.

User	**Cloddy Hans**
Röd fyrkant *Red square*	Nu förstod jag inte *Now I didn't understand*
Röd kvadrat ovanpå blå triangel *Red quadrate on top of blue triangle*	Vill du att jag ska **flytta** den röda **fyrkanten** till den blåa trekanten? *Do you want me to **move** the red **square** to the blue trigon?*
Ja just det *Yes that's right*	Ok *Ok*
Sen ska du **flytta** den blåa cirkeln till den gröna **fyrkanten** *Then you should **move** the blue circle to the green **square***	Nu förstod jag inte *Now I didn't understand*
Flytta den gr_blåa cirkeln till den gröna fyrkanten *Move the gr_blue circle to the green square*	Vill du att jag ska flytta den blåa cirkeln till den gröna fyrkanten? *Do you want me to move the blue circle to the green square?*

In order to investigate whether users adapted their speech at the lexical level, we examined the users' choices of verbs in the first and last turn of their dialogue with Cloddy Hans, respectively. The following verbs occurred in the users' utterances: *flytta* ('move'), *placera* ('place'), *sätt* ('put'), *läggs* ('lay' in the passive voice). In addition, when the verb was omitted entirely, as in the utterance "Blå cirkel till grön fyrkant" ('Blue circle to green square'), the label *no_verb* was used. Finally, fragmented utterances in which one of the references to the objects was missing, as in "Röd fyrkant" ('Red square') were labeled as *incomplete*.

The users' choice of lexical items in the reference to objects in the dialogue also seemed interesting from the point of view of convergence. In particular, we focused on those lexical items that in Swedish have several near-synonyms to see whether subjects would coordinate their references to match those of the systems'. As with the verbs above, we were interested in comparing the speakers' choice of vocabulary in their first and last turns to Cloddy Hans. The system consistently used the term *trekant* ('trigon'), and the users also employed the terms *triangel* ('triangle') and *trehörning* ('tripod'). Similarly, throughout the dialogues the system used the term *fyrkant* ('square'). In addition, the subjects' utterances contained the lexical items *kvadrat* ('quadrate'), *rektangel* ('rectangle'), *ruta* ('box') and *kub* ('cube').

9.2.3 Results

As can be seen in Figure 9.4, users often adapted their choice of verbs to match the one (*flytta,* 'move') occurring in the system's output. Even though it sounded rather strange, some speakers initially used the verb *placera* ('place') because this was the term they had heard in the prerecorded instructions. However, in the last turn of the dialogue not a single occurrence of this verb could be found, indicating that this verb initially was chosen as a result of priming. In the first utterances to the system, an equal number of users chose to avoid using a verb at all, expressing themselves in command-like language, such as "Röd fyrkant till blå cirkel" ('Red square to blue circle'). Slightly fewer speakers used this sort of *no_verb* construction towards the end of their dialogue with Cloddy Hans, but these findings are inconclusive.

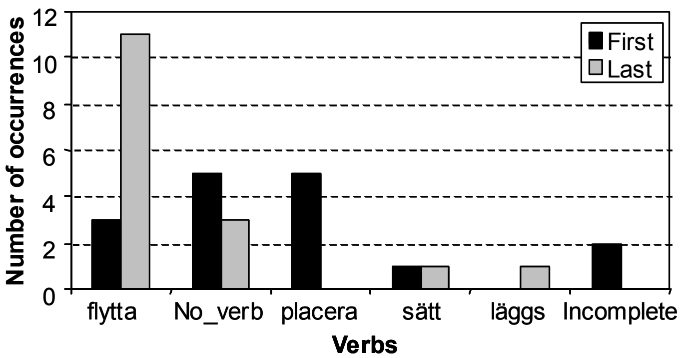


Figure 9.4 The users' choice of verbs in the first and last utterance of their dialogue.

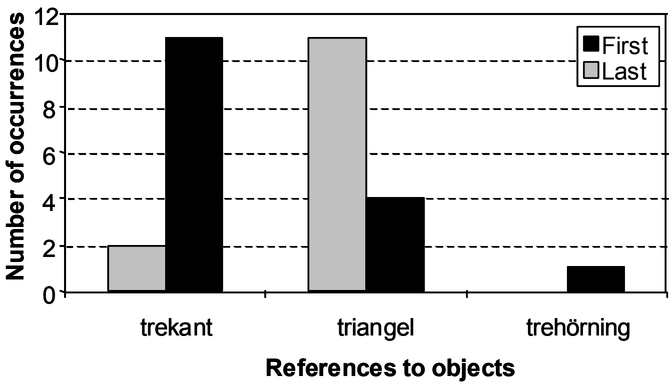


Figure 9.5 The subjects' references to objects in their first and last turns.

As can be seen in Figure 9.5 and Figure 9.6, speakers were strongly influenced by Cloddy Hans' way of referring to objects in the dialogue. It is interesting to note that the system's choice of term for three-legged objects, *trekant*, was not the one spontaneously preferred by the subjects of this experiment as we had anticipated. Instead, Figure 9.5 shows that the speakers initially showed a strong preference for using the word *triangel*, but adapted lexically in the course of their dialogue with Cloddy Hans.

Figure 9.6 is an example of the opposite: the term spontaneously preferred by most users, *fyrkant* ('square'), was also the term which occurred in the system's feedback prompts. However, almost as many subjects initially used the term *kvadrat*. The fact that many subjects used the system's output term from the beginning made the lexical entrainment effect appear less dramatic in this example. Nonetheless, only two speakers used another term than *fyrkant* during their last turn with the Cloddy Hans system.

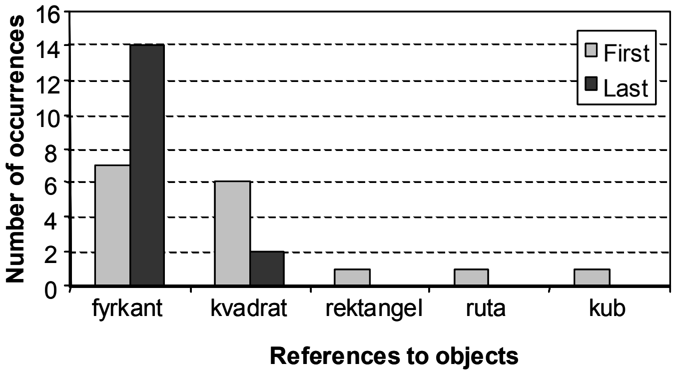


Figure 9.6 The subjects' references to objects in their first and last turns.

9.2.4 Discussion and Summary

The general tendency in the Cloddy Hans corpus, as manifested in the comparisons between the users' first and last utterances to the system, was that the subjects were influenced by the system's lexical choices in the dialogue. As discussed in Chapter 8, post-experimental interviews indicated that subjects were not aware of adapting their speech rate to match that of Cloddy Hans even though they clearly did so. In contrast, these interviews indicated that several of the subjects had intentionally adapted their lexical choices, accepting the terms suggested by Cloddy Hans.

The Wizard-of-Oz-version of the system the users interacted with deliberately 'misunderstood' some of their utterances. Nonetheless, a subject could later on in the dialogue use a lexical item that had previously been misunderstood without encountering a problem. This inconsistency in the system's responses makes it difficult to speculate on which of the users' lexical adaptations might be intentional. A user who had adapted his lexical choices after being misunderstood experienced a few successful turns, and later encountered a problem again. However, because of the performance of current speech recognizers, this is not an altogether unlikely scenario with a fully functional system.

One contributing factor to understanding users' lexical adaptations in the Cloddy Hans corpus was probably the animated agent's general behavior and appearance. The character Cloddy Hans was deliberately designed as to seem a bit slow, with a sluggish look and a wearisome voice. When confronted with such a figure, users may have felt that to imitate him was the only option available to them.

In this study, we examined lexical entrainment effects in the Cloddy Hans database. Results from comparisons of lexical choices in the first and last turns to the system suggested that the subjects adapted to the vocabulary used by the spoken dialogue system.

Chapter 10

Pragmatic Adaptations

In this chapter, we address the question of how users adapt their language at the pragmatic level when they interact with a spoken dialogue system. This chapter includes four corpus studies from two dialogue systems. In the first section of this chapter, we investigate users' socializing behavior during human–computer interaction. In the second, third and fourth sections, we address issues related to user initiative, turn-taking, discourse structure and feedback strategies in the multimodal AdApt corpus.

The first section is based on a series of empirical studies, all of which aimed at increasing our understanding of users' socializing behavior in the August system. Different parts of these corpus studies have previously been published in the *Proceedings of Eurospeech* (Bell and Gustafson 1999b), the *Proceedings of IDS'99* (Bell and Gustafson 1999c) and *Natural Language Engineering* (Gustafson and Bell 2000). While Joakim Gustafson designed, implemented and maintained most of the August system, Linda Bell contributed to the development of the system during the latter part of the project. Both authors contributed equally to all aspects of tagging and labelling the corpus, as well as all aspects of analyzing the results.

In the second study, we discuss turn-taking in the AdApt corpus. The work presented in Section 10.2 is an extended and revised version of the article "Real-time Handling of Fragmented Utterances," which was originally published in the *Proceedings of the NAACL '01 Workshop on Adaptation in Spoken Dialogue* Systems (Bell, Boye and Gustafson 2001). In this study, we present an analysis of fragmented user utterances in the AdApt corpus, performed by Linda Bell and Joakim Gustafson. The results of this analysis led to the development of an incremental interpretation method, implemented by Johan Boye and Joakim Gustafson. The description of this method is not included in the section below. Interested readers are referred to the original article for a complete description.

The third study, a revised and shortened version of an article previously published in the *Proceedings of ICSLP* (International Conference on Spoken Language Processing), discusses the distribution of disfluencies in the multimodal AdApt corpus and a unimodal dialogue corpus collected at Telia Research (Bell, Eklund and Gustafson 2000). Robert Eklund transcribed and annotated the disfluencies in the speech data, and was also responsible for the statistical computations. Linda Bell and Joakim Gustafson were jointly responsible for the multimodal data collection, and the dialogue analyses of both corpora.

The AdApt system did not explicitly ask its users to provide backchannels or feedback. However, during data collection with the AdApt system, it became clear that many users voluntarily provided the system with positive and negative feedback on the state of the dialogue. The final study, describing feedback behavior in the AdApt corpus, is a revised version of an article that was previously published in the *Proceedings of ICSLP* (Bell and Gustafson 2000). The AdApt group at KTH–CTT contributed to the design and set-up of the experimental system. Joakim Gustafson designed and implemented the tools used for collecting the wizard data. Linda Bell and Joakim Gustafson were responsible for the data collection, as well as labelling and analyzing the data.

10.1 Social User Behavior in the August Corpus

One of the aims of the August project, described in Chapter 7 of this thesis, was to collect spontaneous speech input from people who had little previous experience of spoken dialogue systems. When the August corpus was analyzed, it became clear that many of the people who interacted with the system seemed to be more interested in exchanging greetings with the animated agent rather than searching for information. This resulted in large numbers of what in our analysis is referred to as *social* utterances. In addition, the August users often asked factual questions that were clearly out-of-domain, commented on the system itself and previous dialogue turns and sometimes even tried to deceive the system. The August system was designed with a number of simple domains instead of a single complex one, and one of these domains handled greetings and other social utterances. Nonetheless, it was clear that the performance of the system did not always match the users' expectations. This is also reflected in the August corpus, which contains a number of utterances that are referred to below as *insults*.

10.1.1 Background

The concept of illocutionary force (Austin 1961), which was discussed in Chapters 3 and 4, is of central importance in this context. The idea was to categorize the utterances in the August database in accordance with what was presumed to be the users' intentions. In the development of speech interfaces, it is often important to assign speech act labels to user utterances. However, as discussed in Chapter 4 above, there is no one-to-one mapping from the surface form of an utterance to its meaning in a certain context, see for instance Jurafsky and Martin (2000). On the contrary, it can sometimes be very difficult to interpret the illocutionary force of a particular utterance. In this study, we describe the categorization of the spoken input in the August database into utterance types. The aim was to address the issue of how users adapt their language at the pragmatic level. The underlying assumption is that the users' input, and what they try to express through their utterances, is a reflection of their expectations and opinions of the system's capabilities. For example, a user who believes the system capable of being able to answer questions will ask for information, whereas a disappointed user may become ironic or even insult the system. We also briefly examine lexical and syntactic aspects of the August corpus to see whether the utterance types are distinguishable in terms of specific linguistic features or general complexity. Differences between how adults and children interact with the spoken dialogue system are also considered, and implications for future dialogue systems are suggested.

10.1.2 Data and Annotation

The study is based on the August corpus, described in 7.1, and consists of recordings and transcriptions of more than 10,000 utterances of spontaneous computer-directed speech.

In the analysis of social user behavior in the August database, the utterances were labeled according to the presumed *intentions* of the users. The purpose of this categorization was to get a better picture of the kinds of things the users wanted to convey when interacting with the system. Were the users trying to retrieve information or were they merely interested in socializing with the animated agent? As mentioned above, the concept of illocutionary force or communicative intention is a difficult one, both in human–human and human–computer interaction. The categorization of utterances involves an arbitrary element, as one and the same utterance may express different things depending on the context. Moreover, which and how many categories to use can be a problematic issue. After carefully considering these questions, we decided to categorize the utterances in the August database in accordance with a simplified pragmatic model containing six major groups, see Table 10.1 below.

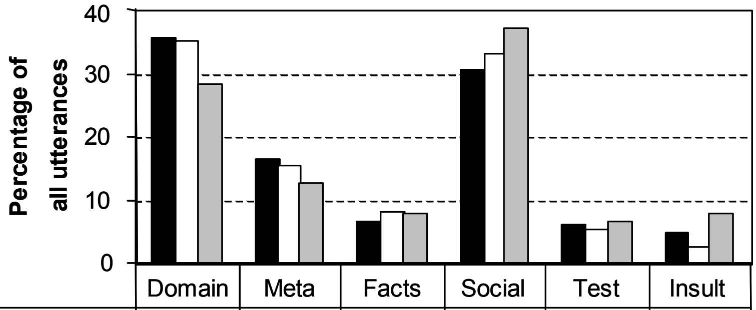
Table 10.1 The utterance types in the August database.

SOCIALIZING	Translated examples	**INFO-SEEKING**	Translated examples
Social	Hello August!	**Domain**	Where can I find restaurants?
Insult	You are stupid!	**Meta**	What can I ask you?
Test	What is my name?	**Facts**	What's the capital of Finland?

The *social* category consisted of greetings and remarks of a personal kind, while expletive expressions and swear words were placed in the category of *insults*. The category called *test* contained utterances that were spoken with what appeared to be the purpose of deceiving the system. The *domain* category included utterances in one of the system's domains, as indicated in the user interface, see Section 7.1. Questions about the system itself and comments about the actual dialogue were grouped into the *meta* category. Factual questions outside the domains mostly turned out to be of an encyclopedic nature and sometimes dealt with things people would expect a computer to be good at, such as calculus. These utterances were categorized as *facts*. These six categories were then brought together into two main groups. The first one, *socializing*, included the categories *social*, *insults* and *test* while the second one, *information-seeking*, included the categories *domain*, *meta* and *facts*.

10.1.3 Results

Many of the August users were interested in socializing with the system, as can be seen in Figure 10.1. Some differences between how adults and children communicated with the system could be observed. Children in the present study made use of social utterances to a greater extent than adults did. One possible explanation might be that the other domains did not particularly appeal to children. In terms of politeness behavior, or the lack of the same, some differences between the groups could also be discerned. As suggested in Figure 10.1, women rarely use insults when interacting with August, while children do it more frequently.



	Domain	Meta	Facts	Social	Test	Insult
■ men	35,6	16,4	6,4	30,8	6,0	4,9
□ women	35,3	15,5	8,2	33,1	5,3	2,6
▒ children	28,3	12,6	7,9	37,1	6,4	7,7

Figure 10.1 Distribution of utterance types in the August corpus.

In order to be able to get a better overview of the utterance categories in the database, we turned our focus to the two main groups *socializing* and *information-seeking*. In the entire corpus, 56% of the utterances were grouped in one of the *information-seeking* categories, and the *socializing* categories constituted the remaining 44% of the utterances. Figure 10.2 below points to the different strategies used by those people whose interaction with the August system lasted for more than two turns.

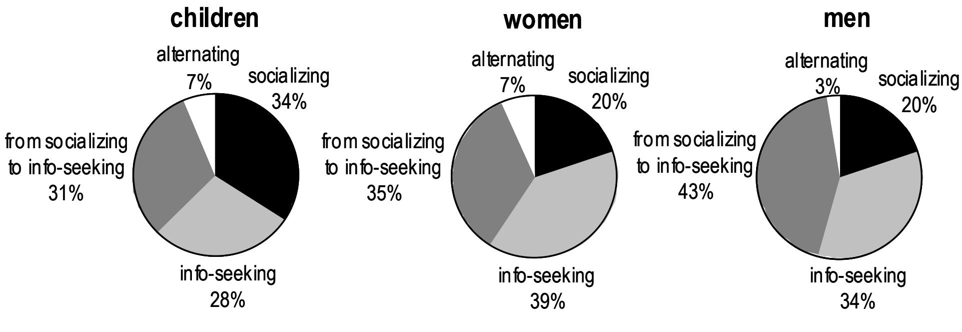


Figure 10.2 The distribution of speakers with respect to their usage of the utterance categories socializing and info-seeking. The statistics are based on the first utterances (up to six) from all users who produced more than two utterances to the system. These constitute 67% of all utterances by children and women, and 58% of all utterances by men.

As indicated in Figure 10.2, men more often began by socializing with the system and then turned the dialogue to the area of information-seeking, while women focussed on more domain- and fact-oriented questions from the beginning. In contrast, many of the children used only social utterances over the first turns. Very few users alternated between information-seeking and socializing during their first six turns. There seemed to be four distinguishable groups of users: first, those who only wanted to socialize, second, those who only wanted to seek information, third, those who began by using some greeting remarks and then turned to information-seeking and remained in that area. The final group was a small one, and it consisted of users who tried to communicate with the system, but failed, and alternated between information-seeking and socializing, trying to get the system to understand.

10.1.4 Discussion and Summary

The analysis of the August database indicated that the users of the August system can be divided into different groups depending on their dialogue strategies. The users seemed to either want to socialize with the system or search for information. An important question was whether it was possible to make the users talk about the system's domains instead of merely socializing. In order to study this in the August corpus, those utterances that occurred immediately before and after certain system prompts were analyzed. These selected system prompts were supposed to be generated when the users had explicitly asked about what they could say to the system. The animated agent then responded either: "I know where certain streets are located" or "I know things about Strindberg, KTH and Stockholm." Figure 10.3 below shows how these prompts influenced the users when they were mistakenly generated due to recognition failure.

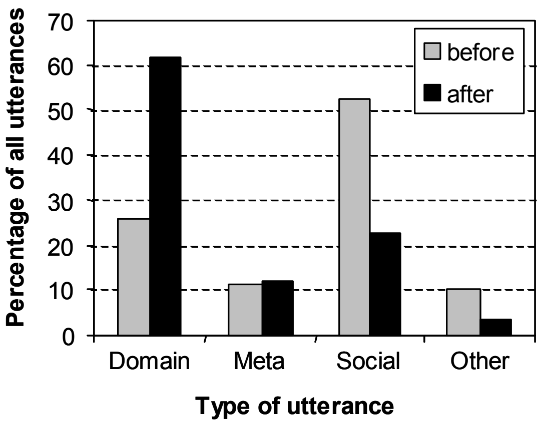


Figure 10.3 The effects of system prompting.

As can be observed, 63% of the users actually conformed to the system by immediately asking about one of the topics mentioned in the preceding prompt. The number of domain-related utterances before these prompts appeared was only 26%. The number of utterances in the socializing category decreased significantly after such a prompt, and only rarely did the users talk about something in one of the other categories.

Considering the fact that the August system had several different domains and that the dialogue model was not strictly specified, the input utterances in the August database were generally quite simple. People with little experience of spoken dialogue systems have different expectations and make use of a variety of strategies as they interact with such a system. Some users started by looking for information immediately, while others preferred to socialize with the system before going into this mode. 44% of all the utterances in the August database were categorized as *socializing*. The animated agent's human-like appearance and the fact that he responded to greetings probably made this social interaction make sense.

In this section, we have addressed pragmatic adaptation in the August corpus by describing our attempts to categorize the database into utterance types. We noted that over 40% of the utterances in the corpus were labeled as *socializing*, and discussed various ways of explaining this observation. The occurrence of insults and meta-language in the corpus either indicates that users believed the animated agent capable of handling such utterances, or that they felt frustrated enough to say anything to this unsupervised system. Children used social language with the system to an even larger extent than adults did, perhaps because their interest in the domains of the system was limited. We also noted that system prompts suggesting topics of conversation influenced users to talk about one of the domains of the system.

10.2 Turn-handling in AdApt

In this section, we discuss how the users of the AdApt system adapt their turn-taking behavior during their human–computer interaction. More specifically, we will discuss how the structure of the dialogue affected user behavior, resulting in fragmented turns.

10.2.1 Background

Even though we may not be aware of it as we engage in conversation, human dialogue is full of hesitations, restarts, corrections and pauses. In contrast to most written language genres, where the language structure and syntax are deliberately planned and organized, spontaneous speech is produced in real time and cannot be edited (Miller and Weinert 1998). Heeman (1997) observes that a speaker's turn in spontaneous conversation is often used to make several contributions. Many turns in spoken dialogue contain fragments, divided by silent pauses:

> "I would like a /pause/ three-room apartment in this area"
>
> "The red apartment /pause/ how much does it cost"
>
> ”Mhm that seems good /pause/ when was the building constructed"

As discussed in Chapter 3 above, the seemingly effortless process of turn-taking in human dialogue is governed by a set of rules. By analyzing spontaneous human–human dialogues, Sacks et al. (1974) discovered that there are certain places in the dialogue where it is possible for a participant to take over the floor from the person presently speaking. These points in the dialogue, *transition-relevance places*, are marked with linguistic as well as non-linguistic cues (Levinson 1983). For example, prosodic markers and eye-gazing signals can be helpful for a listener wishing to detect a possible place for transition.

When one of the dialogue partners is a computer, the issue of fragmented speech input and turn-taking becomes more complicated. While advanced spoken dialogue systems elicit fragmented turns from their users, the natural language processing modules of current systems lack the ability to accurately predict a transition-relevance place. Partly, the users' behavior can be explained by the fact that when human–computer dialogues appear more human-like, user expectations are raised correspondingly and a number of human discourse features are introduced into the computer-directed speech. Furthermore, spoken dialogue systems with an open-microphone speech recognizer (rather than click-to-talk) make it more difficult for users to plan their utterances 'off-line' before speaking. Instead, users might begin to speak and plan their next utterance before they are certain of what they want to accomplish. In addition, a system that produces multimodal output with graphical information concurrent with speech will impose a heavy cognitive load on its users. Users of such systems will need more time to consider the information they are presented with, something which may further magnify this behavior. Finally, a dialogue system that hands over the initiative to the user opens up for a greater variability in input responses. This means that users are more likely to hesitate (as in the example utterances above) than if the system keeps the initiative to itself.

The question addressed in this section is how fragmented user turns to a dialogue system really should be interpreted and handled. The intuitive answer, as we understand it, is first of all that a system should be flexible enough to choose from several possible reactions when it detects that the user is silent. More specifically, it should either:

(1) start producing a response, or
(2) give no reaction at all, as more input seems likely to come, or
(3) produce some kind of back-channelling reaction, encouraging the user to continue speaking.

Empirical studies of the AdApt corpus indicated that it would be necessary to incrementally interpret the users' input in real time, and determine the system's reaction at every silent pause. As the appropriate choice of system reaction turns out to be highly dependent on the dialogue state, the system continuously adapts this interpretation process to the current state of the dialogue[5].

Some previous studies have emphasized the importance of incremental interpretation of user input. Allen, Ferguson and Stent (2001) argue that incremental interpretation of user input and flexible turn-taking is necessary for the interaction with spoken dialogue systems to become more natural. Arguments for an asynchronous dialogue model have previously been put forward by Boye, Hockey and Rayner (2000). Nakano et al. (1999) describe a combined parsing and discourse processing method, where the user's utterance can be interpreted each time a word hypothesis comes in from the speech recognizer. Their system interprets user input incrementally in real-time, but does not take the dialogue context into account.

Developers of spoken dialogue systems have traditionally assumed that a silence of a certain length indicates that the system should take the floor. The user has thereby been charged with the task of producing unbroken, continuous spoken input. However, state-of-the art commercial speech recognizers support a method for adapting the end-of-speech detection to the

[5] A detailed description of the incremental interpretation can be found in Bell, Boye and Gustafson (2001).

speech recognition grammar, see for instance the Nuance Guide (2000). The purpose of this is to make the system seem more responsive and 'alert'. For example, a short pause is sufficient to signal end-of-speech if the last word recognized is identified as the last word of an utterance as defined in the grammar. On the other hand, if the last word recognized is not defined as an end-of-utterance word, the system will wait longer before signaling end-of-speech. This method is only applicable in dialogue contexts where the users' responses are to a large extent predictable and a strict grammar can be used. However, the AdApt system uses a statistical grammar based on collected data which makes it difficult to use this sort of feature.

10.2.2 Data and Annotation

Details concerning the AdApt system and data collection can be found in 7.2. above. A Wizard-of-Oz experiment was set-up to collect data for the AdApt corpus, see Chapter 7 (Bell, Boye, Gustafson and Wirén 2000). A total of 32 dialogues with 16 subjects were manually labeled for the purposes of the current investigation.

An example of a dialogue between the animated agent Urban and a user could be seen in 7.2. In Table 10.2 and Table 10.3 below, we present examples of how the system and user utterances were tagged, respectively. The dialogue act tags in Table 10.2 represent the categories of the system's output, as selected by the wizard of the system. These dialogue acts were realized through verbal and graphical output, as seen in the example column in the middle. The column on the right represents the system's degree of initiative, and/or its ability to execute the user's previous request. This categorization was added because it influenced the way the users constructed their next turn.

Table 10.2 Tagging of system output in AdApt.

Dialogue Act	**Example**	**Initiative**
INTRO	Hej, jag heter Urban och kan berätta om fina bostäder i Stockholms innerstad *Hello, my name is Urban and I can tell you abut nice apartments downtown Stockholm*	System turned over initiative
INTRODUCE-GROUP-OF-APARTMENTS	I det markerade området hittar jag 4 lägenheter /fyra lgh-ikoner visas på kartan/ *In the shaded area I find 4 apartments* */four apt-icons shown on map/*	System turned over initiative
INTRODUCE-INDIVIDUAL-APARTMENTS	Det finns en lägenhet på den här gatan, den visas nu på kartan /en lgh-ikon visas på kartan/ *There is an apartment on this street, it is now shown on the map /an apt-icon shown on map/*	System turned over initiative
INFO-INITIATIVE	Vill du veta mer om lägenheten? *Would you like to know more about the apartment?*	System turned over initiative
ASK-FOR-PREFERENCE	Hur många rum ska lägenheten ha? *How many rooms do you want?*	System asked question
TOO-MUCH-TO-PRESENT	Det finns för många sådana lägenheter för att visa. Finns det något särskilt du vill att lägenheten ska ha? *There are too many such apartments to present. Is there anything special you would like your apartment to have?*	System asked question
ANSWER-WITH-INFO	Det vita huset är byggt 1906 *The white building was constructed in 1906*	System answered
ANSWER-WITH-NO-INFO	Det har jag inga uppgifter om *I have no information about that*	System failed to answer
NOTHING-TO-PRESENT	Det finns tyvärr inga sådana lägenheter i det här området *Unfortunately there are no such apartments in this area*	System failed to answer
ASK-FOR-CLARIFICATION	Jag förstår inte vad du menar *I don't understand what you mean*	System failed to answer

Table 10.3 Tagging of user input in AdApt.

Dialogue Act	Example
PREFERENCE	Jag vill att huset ska vara byggt på 1900-talet *I would like the building to be constructed in the 20th century*
ASK-FOR-INFO	Har den svarta lägenheten parkettgolv *Does the black apartment have a tiled floor*
META	Vad menar du med det *What do you mean*
SOCIAL	Hej Urban *Hello Urban*
HOLD-THE-FLOOR	Ja vänta *Yes wait*
FEEDBACK	Mhm det verkar bra *Mhm that seems good*
SELECT-APARTMENT	Den röda lägenheten *The red apartment*

The first two dialogue act categories in Table 10.3 (*preference* and *ask-for-info*) represent domain-related user input. About 96% of all user turns contained one of these dialogue acts. The dialogue acts labeled as *meta* are mostly clarification questions about the system's previous output. In contrast to the August corpus, see 10.1, the AdApt corpus contained very few *social* remarks, probably because the users were focused on solving a rather complicated task.

A user turn in the AdApt corpus often contained more than one dialogue act. The last three dialogue acts (*hold-the-floor*, *feedback* and *select-apartment*) very rarely occurred in turns of their own. Instead, they were used to take the turn before users moved on to a domain-related request. The distribution of these dialogue acts in the AdApt corpus is shown in Figure 10.4. As can be seen, this turn-taking strategy commonly occurs when the system turned over the initiative to the user or explicitly asked for information.

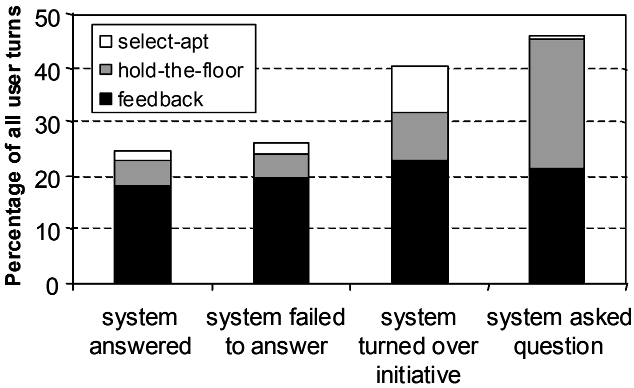


Figure 10.4 Distribution of dialogue acts as initial parts of user turns, depending on the system's previous turn.

User input categorized as *hold-the-floor* consisted of phrases like "wait a minute" or initial filled pauses (such as "eeh" and "mhm"). The frequency and distribution of filled pauses and other speech disfluencies in the AdApt corpus are discussed in Section 10.3. The analysis of user *feedback*, described in detail in Section 10.4, showed that the interpretation of these cues was highly dependent on the dialogue context. As reported in this section, as many as 18% of all user turns in the corpus contained positive or negative feedback.. However, the feedback cues occurred in a separate turn in no more than 6% of all cases.

A turn in the corpus is a sequence of fragments $F_1 \ldots F_n$, divided by silent pauses. Each turn was given *n* tags reflecting its status, at every pause, in the above regard. Two types of status tags were used. The tag *non-closing* meant that the input up to that point could not reasonably be considered as complete, and that the system should not yet begin preparing its response. All other fragment sequences were labeled as *closing*. In the course of what was in fact a single user turn to the system, several tags of each type were sometimes assigned. For example:

F_1	F_2	F_3	F_4
I would like a …	three room apartment …	on Södermalm or…	Vasastan…
non-closing	*closing*	*non-closing*	*closing*

As mentioned above, a silent pause and an additional request for information followed most of the feedback cues in the dialogues. This example shows a typical dialogue excerpt:

System: In the area marked on the screen I found five apartments (*graphical information presented on the screen*)
User: **Yes** /silence/ ehh is there one with a stuccoed ceiling?

Here, the user explicitly acknowledges that he has understood the system's output. The anaphoric pronoun 'one' refers back to the antecedent 'five apartments' in the system's previous turn.

A frequently occurring pattern in the database was a user utterance that contained a reference to an object shown on the screen, followed by a silent pause. These types of utterances can be given either one of two possible interpretations. In some dialogue contexts, it would seem reasonable for the system to wait for additional information from the user after the first fragment:

System: I found seven apartments and will now display their locations on the map (*colored icons appear on the screen and corresponding information with addresses in a table*)
User: Hagagatan 14 /silence/ when was the apartment built?

At this stage in the dialogue, it would be difficult to come up with a useful interpretation of the user's referential expression. In other dialogue contexts, however, the referring expression supplies the system with sufficient information in itself:

System: This one has a balcony (*highlights red icon*)

User: The yellow one?

In the example above, the user's verbal reference to an apartment icon displayed on the screen in conjunction with the fact that the feature "balcony" was mentioned in the previous turn is enough for the system to be able to fill in what is presupposed. The interpretation of the second user utterance can then be spelled out as: "Does the yellow apartment have a balcony?"

Depending on the dialogue context, one and the same referential expression can either be classified either as *non-closing*, which implies that there is more to come, or as *closing,* which means that the fragment can be given an elliptic interpretation, and contains enough information in itself.

10.2.3 Results

The categorization of the AdApt corpus showed that about 60% of all turns contained a single closing fragment, while 8% contained closing fragments that were followed by at least another fragment. However, as many as one third of the turns in the corpus were labeled as containing a non-closing fragment followed by a silent pause and one or more additional fragments.

The average pause length, for closing as well as non-closing fragments, was about one second. Nine out of ten silent pauses after a closing fragment sequence were 2.5 seconds or shorter. The corresponding figure for non-closing fragments was 3.5 seconds. As can be seen in Figure 10.5, the pause lengths in certain categories of non-closing user input tended to be longer than in others.

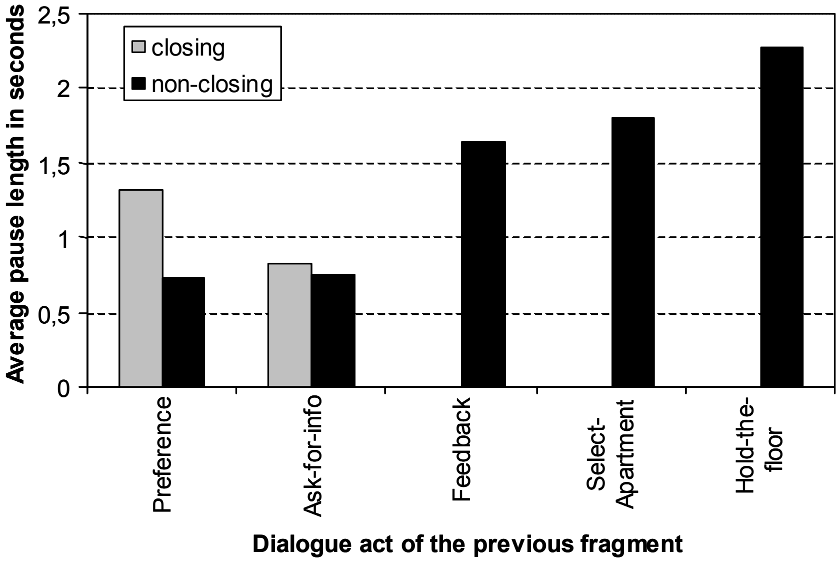


Figure 10.5 Pause lengths after closing and non-closing fragments in different categories of user input

Pauses after non-closing fragments in the dialogue act categories *preference* and *ask-for-info* were on the whole quite short, as seen in Figure 10.5. These pauses were mostly hesitations in the midst of requests for information, as in the example "I'd like a two-room apartment in the Old Town or /pause/ downtown". Most of these fragments could be identified as *non-closing* by the system's parser.

As can be seen in Figure 10.5, the pauses after the non-closing fragments in the dialogue act categories *feedback*, *select-apartment* and *hold-the-floor* were on average longer. In these cases, the users' initial pauses (before uttering anything at all) were also rather extended, which implies they had to take time to consider the system's previous turn.

Certain dialogue contexts appeared to more frequently elicit fragmented user input with inserted silences than others. Long pauses after non-closing fragments were elicited when the system had presented the user with information, graphically and verbally, and then handed over the initiative. In the cases where the system handed over the initiative to the user (*system turned over initiative*), either by saying it had nothing to present or by explicitly asking the user: "What else do you want to know about the apartment?," more than half of all user responses were found to contain a non-closing fragment sequence. However, most non-closing occurrences (63%) appeared after the system had answered a question about some feature of a specific apartment. Most of these non-closing fragments were positive or negative feedback cues, as described in

Section 10.4. In 12% of all cases, the non-closing occurrences appeared in contexts where the system had found a number of apartments that matched the preferences of the user, and presented these options both graphically and verbally. In these cases, it is possible that the system's output increased the user's cognitive load to the extent that hesitation pauses were necessary.

10.2.4 Discussion and Summary

In this section, we have discussed how users adapt their turn-taking behavior in the multimodal AdApt system. Empirical studies showed that the users' input to the system frequently consisted of fragments, divided by silent pauses of various lengths. The system's output and the users' input was categorized into dialogue acts, so that the user's turn-taking strategies could be better understood. It was observed that certain system turns resulted in a frequent use of filled pauses, floor-holders and feedback.

Fragments were categorized as *closing* or *non-closing*, depending on whether the system could start processing the user's input or not. Results showed that one third of the utterances contained a non-closing fragment followed by a silent pause and at least one additional fragment. Certain dialogue contexts more frequently elicited non-closing fragments. In particular, these fragments were produced after the system had presented the user with a lot of information, graphically and verbally and handed over the initiative to the user. Future work includes exploring methods of improving end-of-turn detection in real-time by using a combination of acoustic, lexical and discourse context cues.

10.3 Disfluencies in Two Dialogue Corpora

In the four dialogue corpora described in this thesis, we saw that the different scenarios and tasks resulted in substantial variabilities in the spoken language input. For instance, speech disfluencies were relatively infrequent in the August and Pixie corpora.[6] Part of the explanation for this can be found in the average utterance length in these corpora, which was rather short. The absence of disfluencies can also be explained by the fact that the users of these systems did not have a complicated task to solve. In the Cloddy Hans corpus, disfluencies were also rather infrequent. Those that did occur were mostly found in the users' initial turns, before they knew how to address the system. In contrast, many users of the AdApt system were disfluent throughout their dialogues. In particular, there were certain dialogue contexts that appeared to elicit hesitations, filled pauses, repairs and other disfluencies.

This section compares the frequency and distribution of disfluencies in two spoken dialogue corpora. The AdApt corpus, which was collected by asking subjects to solve a complicated task by interacting face-to-face with an animated agent, was compared to a unimodal travel-booking corpus collected over the telephone.

10.3.1 Background

Spontaneous spoken language contains disfluencies (pauses, truncations, prolongations, repetitions, false starts etc.). For spoken dialogue system applications, disfluencies can be problematic, since current automatic speech recognition is limited in its ability to process them. Depending on the type of discourse or task involved, the type and frequency characteristics of disfluencies will vary.

Scenario and task details affect disfluency rates, frequency and distribution. Previous studies have shown that long, spontaneous utterances tend to be more disfluent than brief, structured interaction (Shriberg 1994; Oviatt 1995). Individual predispositions are also important. Research has shown that some speakers are consistently more disfluent than others (Shriberg 1994; Branigan, Lickley and McKelvie 1999). Other factors, such as planning difficulties, speech rate, confidence, social relationships and gender have also been discussed in conjunction with disfluencies (Shriberg 1994; Bortfeld et al. 1999). User expectations and previous experience with spoken dialogue systems might also play a role.

---

[6] Disfluency rates for the August and Pixie corpora were in the order of 1-1.5% at the word level. However, there was a tendency in these data that children among the users were more disfluent.

In a study where multimodal interaction was compared with a unimodal system that supported speech alone, Oviatt (1995) reported that multimodal interaction tended to contain briefer and simpler language. Multimodal interaction has also been shown to be advantageous from the point of view of error handling, since users tend to switch from one modality to another when their interaction with the computer becomes problematic (Oviatt and VanGent 1996).

10.3.2 Data and Annotation

This study is based on a comparison between two Swedish corpora of human–machine interaction, briefly described below. The first corpus is a single-channel travel booking corpus, which was collected at Telia Research. This database is not included among the corpora described in Chapter 7, but is discussed here for the sake of comparison. The other set of data is the AdApt corpus collected at KTH–CTT, previously described in 7.2.

Unimodal/Human–Machine

This corpus contains human–machine business travel booking dialogues, collected over the telephone. A human operator functioning as wizard simulated the system's speech recognition, while all other components were authentic. The corpus consists of 16 speakers (9 male, 7 female). The subjects were all Telia employees, and were used to the task of booking business trips. In order to avoid linguistic bias, the subjects were given the tasks in pictorial form, and they were also given some time to prepare the task. All subjects believed they were talking to a functional system.

Multimodal/Human–Machine

The multimodal AdApt corpus, described in Chapter 7, contains speech and graphical data from users who interacted with a semi-simulated multimodal dialogue system (Bell et al. 2000; Gustafson et al. 2000). The corpus consists of 16 speakers (8 male, 8 female), each of whom performed two dialogues with a Wizard-of-Oz version of the system. Subjects were informed that they could use either speech or graphical input at any time during the dialogues. Post-experimental interviews showed that all users had been unaware of the fact that this was not a real system.

The corpora were labeled for disfluencies according to an annotation scheme described in Eklund (1999). This system is based on the annotation scheme developed by Shriberg (1994), with some extensions and minor changes. In addition to the disfluency tags, both corpora were labeled with utterance types roughly corresponding to the illocutionary force of the user's utterance. The following labels were used:

Initial - the initial turn of each dialogue

Preference - preferences like destination and number of rooms

Ask - question within the task

Meta - question about the system itself, or its capabilities

Change - changing of features such as departure time of a suggested trip

Repeat - the user asks for repetition (only in the multimodal corpus)

Social - greeting (only in the multimodal corpus)

10.3.3 Results

The average number of disfluencies per word was found to be 12.2 % in the unimodal corpus and 5.4 % in the multimodal corpus. These figures exclude unfilled pauses and utterance-initial filled pauses, which were not labeled as disfluencies in the present study. As can be seen in Figure 10.6, there is a difference between the two corpora with regard to the number of disfluencies as a function of utterance length.

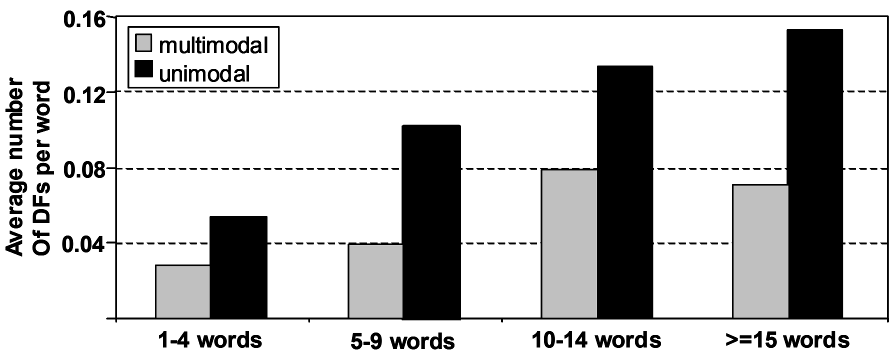


Figure 10.6 Number of disfluencies per word as a function of utterance length.

In the unimodal corpus, the figures are in line with previously reported studies in that the disfluency rate increases as a more or less linear function of utterance length. The multimodal corpus deviates from the norm by displaying fewer disfluencies in the utterances that were more than 15 words than those that were 10-14 words. As will be shown below, this can partly be attributed to the speech act function of the utterances in which the disfluencies occur, see Figure 10.8.

According to Shriberg's (1994) report on individual 'styles' of disfluency, certain speakers are more likely to use repetitions while other speakers exhibit a relatively high number of deletions. Furthermore, Branigan et al. (1999) show that frequent occurrences of one type of disfluency for an individual speaker often correlate with high frequencies of another type of disfluency. Thus, some speakers seem to be more disfluent than others, regardless of the type of disfluency. In the present study, a few of the speakers in both the unimodal and multimodal corpus exhibited a strikingly high number of disfluencies, relatively speaking. As can be seen in Figure 10.7, these individual differences are apparent even in turns of average length. There are even two speakers in the unimodal corpus and two speakers in the multimodal corpus who were not disfluent at all. Individual variation thus exceeds most other kinds of factors in explaining disfluency rates. In our data, factors such as gender, age and computer skill had no effect on disfluency rates.

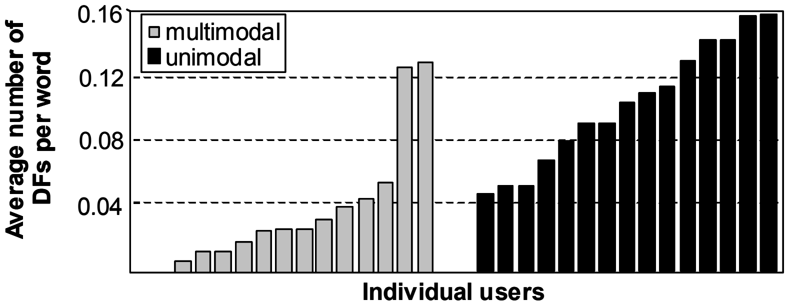


Figure 10.7 The average disfluency/word rates for turns with five to nine words.

According to Oviatt (1995), the structure of the dialogue affects the manner in which users interact with a spoken or multimodal dialogue system. A system that employs an unconstrained format will encourage its users to produce utterances with higher information-per-utterance ratio than users who are prompted for more specific information. In the unimodal corpus, the system greeted the subject with an open question like "Welcome to the travelling service. How may I help you?," while the opening utterance from the multimodal system was the more constraining: "Hello my name is Urban. I can help you find apartments in Stockholm. Where would you like to live?" This could explain that the average length of the first user turn in the unimodal corpus is 17 words, while the first user turn in the multimodal corpus is 10 words on average.

Another factor which is likely to have affected the collected data is that the wizard of the unimodal system was not explicitly instructed to limit the number of words in an utterance that he should 'understand', nor was he instructed to misunderstand otherwise problematic utterances. Similarly, the wizard of the multimodal system 'understood' long and fragmented turns within the domain of the system. However, this wizard did not 'understand' user input with complicated syntax or out-of-domain words.

As is indicated in Figure 10.8, disfluency rates in the unimodal corpus were highly dependent on the utterance type in the dialogue in which they occur. In some cases, the utterance type appears to be more influential than utterance length as a way of explaining disfluency distribution.

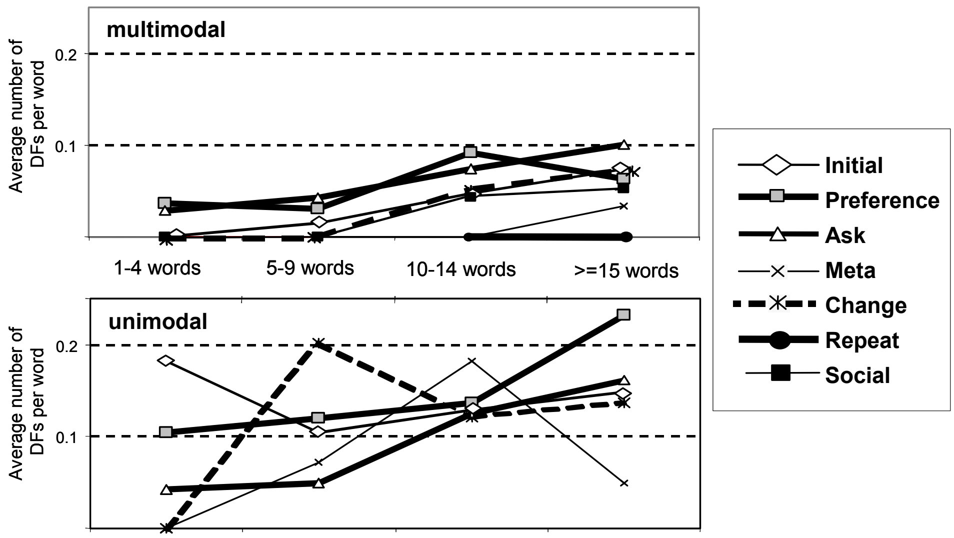


Figure 10.8 Number of disfluencies per word as a function of utterance type.

In the multimodal corpus, utterance type does not seem to affect disfluency distribution in a significant way. However, in ten-word sentences or longer, there is an increase in disfluency production irrespective of utterance type. In the unimodal corpus, it was clear that certain stages in the dialogues required a lot of planning on the part of the user. In particular, this was the case when the users were asked to specify their departure times. The scenarios specified scheduled times for meetings, conferences etc., and the users had to figure out for themselves when they had to arrive at the destination in order to make it on time for their appointment. Naturally, this required more planning and effort than the simple 'slot-filling' questions that were frequent at other places in the dialogues. Consequently, the peaks for 'preference' for the unimodal corpus in Figure 10.8 can be explained by the elevated disfluency figures for these specific turns. When the users of the unimodal system suggested a departure time, the system sometimes proposed, in detail, a trip with too late an arrival time for the user to be able to make his or her appointment. This led subjects to enter into a clarification subdialogue with the system, and these attempts to negotiate with the system often yielded long and highly disfluent user utterances. These utterances are labeled as 'change' in Figure 10.8. This tendency could also be seen in the multimodal system, albeit to a lesser extent.

At certain points in the unimodal as well as in the multimodal dialogues, open questions from the system could be assumed to have encouraged the subjects to express themselves with some verbosity. The average utterance length in both corpora was about seven words. However, after an open question from the system the average utterance length in the unimodal system increased to 13 words, while the corresponding figure for the multimodal corpus was 9.5 words. Thus, the tendency to become more verbose after unconstrained questions appears to be more accentuated in the unimodal corpus. The reason for this is probably that the open questions in the multimodal system were within the task at hand, while the open questions in the unimodal system initiated a new (sub-)task. In the multimodal system a typical open question was "What else do you want to know about the apartment," while a typical open question in the unimodal system was "I have booked a flight from *A* departing at *T1* to *B* arriving at *T2*. What else do you want to book?" It is likely that the subjects were affected by the system's verbose summary of the booking. Since the system 'understood' long and informationally dense utterances from the start, the users may have been implicitly encouraged to supply the system with as much information as possible in a single turn.

At the grammatical level, one notable difference between the corpora is the occurrence of topicalized utterances in the multimodal corpus, that are not found in the unimodal corpus. A total number of 28 utterances in the multimodal corpus have the form "Den gröna fastigheten, har den balkong?" ("The green building, does it have a balcony"), rather than the standard "Har den gröna fastigheten balkong?" ("Does the green building have a balcony"). These topicalized sentences are characteristic in that the fronted item is followed by either a filled pause or an unfilled pause, e.g., "Eh den röda fastigheten på Swedenborgsgatan, eh har den balkong?" ("Eh the red house on Swedenborgsgatan, eh does it have a balcony?"). The fact that the users of the multimodal system have the discourse objects visually available, at least during certain stages of the interaction, seemingly has an effect on both the grammar and the disfluency distribution.

10.3.4 Discussion and Summary

A number of factors contributed to the differences in disfluency rates between the two corpora. The scenarios were not identical, and the time-planning feature of the unimodal dialogues can be assumed to have influenced the results significantly. There was a greater number of very long sentences in the unimodal corpus, which raised the disfluency rates in this corpus. These, and probably other factors, contribute to the differences in results reported for the corpora. However, some of the observed dissimilarities can be ascribed to the modality used in the collection. Oviatt (1995) reports that human–human telephone speech is more disfluent than face-to-face conversations. Similar results for human–computer interaction were reported by (Bickmore and Cassell, forthcoming). This could explain the overall higher disfluency rate in the unimodal corpus as compared to the multimodal corpus. Adding a face seems to increase the naturalness of the interaction. Despite the fact that the animated face in the multimodal system was not a real human face, the multimodal corpus contains a higher degree of social and conversational behavior than the unimodal corpus.

A clear difference in the interface modality dimension is that a telephone interface puts heavier demands on the buffer memory of the user when the system presents information than does a graphical interface. This could explain the higher frequency of disfluencies in the unimodal corpus in interactional stages where the user has to react to information output from the system, while at the same time keeping, and accessing, the required information in their working memories. As has been shown, the occurrence of topicalized utterances in the multimodal corpus shows that the way information is presented to the user affects the syntax of the users' responses, and consequently also disfluency distribution.

In this study, we compared the distribution of disfluencies in two human–computer dialogue corpora. One corpus consisted of unimodal travel booking dialogues, which were recorded over the telephone. The other corpus was AdApt, which was collected using a simulated version of a multimodal dialogue system with an animated talking agent. The results showed that the unimodal corpus contained a higher number of disfluencies. Furthermore, it appeared as if certain utterance types in the unimodal corpus and, to a lesser extent, in the multimodal corpus, elicited user disfluencies. We discussed these findings and related them to possible effects of modality, task and interface design.

10.4 Feedback in the AdApt Corpus

Acknowledgements and feedback cues are characteristic features of human–human conversation. As discussed in Chapter 4 above, early studies of human–computer interaction indicated that people would be unwilling to provide a dialogue system with backchannels of this sort (Reilly 1987). However, investigations of the AdApt corpus offered evidence to the contrary. In this study, we explore pragmatic adaptation as it is manifested in users' positive and negative feedback to the AdApt system.

10.4.1 Background

As conversational speech interfaces become more advanced and human–computer dialogues appear more natural, we may expect users of spoken dialogue systems to integrate a larger number of human discourse features into their speech. In human–human conversation, dialogue participants continuously give each other positive and negative feedback as a way of showing attention, recognizing the intention of what the other conversant is saying or to signal non-understanding or misunderstanding (Allwood 2000). In the present study, we examine a broad range of feedback phenomena observed in a spoken dialogue corpus. The AdApt system is designed to provide users with information about apartments in downtown Stockholm, and for the purposes of the present study a semi-simulated version of the system was employed. Despite the fact that this system never gave the subjects any explicit acknowledgements in the course of the dialogues, positive and negative feedback occurs in a surprisingly large number of user turns.

Clark's theory of grounding, discussed in Chapter 3 above, describes discourse as a joint activity in which participants continuously work at establishing a common ground (Clark 1996). According to this theory, dialogue participants use acknowledgements and feedback to signal understanding and non-understanding throughout the conversation. These cues often carry important information about the grounding process and the state of the dialogue. As mentioned in Chapter 3, Clark and Schaefer (1989) suggest that there are a number of ways in which a dialogue participant can demonstrate that he has understood a discourse contribution. "Acknowledgement" is placed in the middle of a scale ranging from "continued attention" to "display." In dialogue, an acknowledgement is often expressed by a nod or a "yeah," "uh huh" or something similar. Brennan and Hulteen (1995) present a list of acknowledgement strategies that is partly based on Clark and Schaefers' scale, and emphasize the importance of feedback for coordinating the user and systems' knowledge states in a dialogue system and for facilitating problem solving (Brennan and Hulteen 1995).

As reported in Chapter 5 above, Ward and Heeman (2000) found that acknowledgements were frequently used in a telephone-based automated service system. Even though this system did not explicitly encourage the use of feedback, it provided opportunities for and responded to acknowledgements. In this study, about half of the subjects were found to use acknowledgements at least once during their interaction.

Allwood and colleagues (Allwood, Nivre and Ahlsén 1993; Allwood 2000) propose a model of communication and pragmatics which includes a description of different types of speaker feedback in human dialogue. In this framework, linguistic feedback enables speakers to exchange information about basic communicative functions. Feedback utterances are labeled as either positive or negative, and as either explicit or implicit (Allwood 2000). Although independently developed, the scheme used to analyze user feedback in the AdApt corpus of human–computer interaction uses a similar categorization, see Section 10.4.2.

10.4.2 Data and Annotation

AdApt, described in 7.2, is a Swedish conversational multimodal dialogue system which can be used for accessing information about apartments for sale in downtown Stockholm. A Wizard-of-Oz experiment was set-up to collect data for the AdApt corpus, see Chapter 7 (Bell, Boye, Gustafson and Wirén 2000). In addition, for the purposes of the present study a corpus using a modified version of the AdApt wizard interface was also analyzed (Skantze 2002). All in all, 50 dialogues with 33 subjects and a total of 1,845 utterances were tagged and labeled for feedback.

A spoken dialogue system's way of providing feedback affects the users' manner of interacting with that system. The simulated version of the AdApt system did not explicitly acknowledge that the subjects' input to the system was being processed or had been correctly recognized. However, indirect visual cues were conveyed through the system's animated talking head. While speech input was being processed, the talking head appeared to be "listening," and as soon as a user had finished speaking, the head indicated that the spoken input was being interpreted by responding with a "thinking" gesture. Furthermore, by appearing to understand most of what was being said, the system's wizard could be said to have indirectly encouraged the subjects' conversational behavior. In the course of the dialogues, the system continuously offered implicit evidence of understanding. An example from the AdApt database illustrates this:

Table 10.4 An excerpt of a dialogue from the AdApt corpus.

System	User
Var i Stockholm vill du bo? *Where in Stockholm do you want to live?*	Jag vill bo i Gamla stan *I want to live in the Old Town*
Hur många rum vill du ha? *How many rooms do you want?*	Två eller tre *Two or three*

In the example in Table 10.4, the system indirectly acknowledges the subject's input by continuing with a relevant dialogue turn without requesting a repetition of the user's previous utterance. A few turns later, when the system has found a selection of apartments in the Old Town and they are displayed on the screen, the user will know for certain that this turn was correctly interpreted. If the system had used an explicit acknowledgement strategy instead, the system's response to User 1 would for example have been: "The Old Town. Is that correct." If this sort of explicit prompt had been employed, user feedback strategies in the AdApt corpus would probably have been different. As discussed in Chapter 5, intermediate strategies where the system's acknowledgement is part of the next turn, are also a possibility.

The corpus of 1,845 utterances was manually transcribed and the subjects' utterances were individually labeled for feedback, taking into account the context of the system's previous utterance and the dialogue history. For example, when "no" was used as a way of signalling dissatisfaction or disagreement in the dialogue, it was marked as feedback. Conversely, when "no" occurred as response to a question posed by the system, it was not labeled as feedback. Those parts of the user utterances that had been marked as feedback were then tagged with respect to the following three parameters:

Positive/Negative

Positive feedback typically includes expressions like "good," "yes," and "thank you." Examples from the *negative* feedback category include "no," "well" and "too bad." Since some expressions, such as "okay," function as either a positive or negative cue, all sound files were individually assessed. Prosodic or contextual cues indicated whether an utterance was intended by the subject as a positive or negative response to the system's previous utterance.

Explicit/Implicit

In some of the feedback utterances the subjects literally expressed what they meant, so that for example a presentation of a new apartment would get the response "that's great" or "very good Urban." These were labeled as *explicit*, while those utterances where the feedback was conveyed in a less direct way were labeled as *implicit*. Implicit feedback was often expressed through cues like "mhm," and "aha, all right." Again, some cases were ambiguous.

Attention/Attitude

Attention was interpreted as an indication from the user that the system's message has been received. Typical examples include "I see," and "No bathtub." *Attitude*, on the other hand, was seen as an indication of the user's attitude toward the system or the previous turn in the dialogue. Positive and negative value judgements occur frequently in this category. Examples from the corpus include "that's good," "great, Urban," "thanks," "that was quite expensive" and "too bad."

All feedback utterances were categorized along these three axes, resulting in a total of eight groups. As previously observed, some expressions in the corpus turned out to be inherently ambiguous. The word "okay," for instance, was labeled as belonging to all of the categories depending on the context in which it appeared. Table 10.5 shows part of an annotated dialogue sequence in which examples of most of these labeling categories are included. In this excerpt, the user gives the system feedback at every turn. Most of the feedback was labeled as positive. The single instance of negative feedback from the user, turn 36 in Table 10.5, is a response to a system turn that conveyed no new information.

Table 10.5 A translated excerpt from the AdApt corpus. The part of the user utterance that has been labeled as feedback is in boldface, and the type of feedback — positive/negative, explicit/implicit, attention/attitude — is to the right.

#	System	User	Feedback		
32	This building was constructed in 1680	**Yes that is very old indeed** ....when did you say the white one was built	Pos	Exp	Attitude
33	This building was constructed in 1861	**Yes yes that's right**.....is there a tiled stove there too	Pos	Exp	Attention
34	This apartment has a fireplace	**Yes that's all right too** …….. how high is the building	Pos	Exp	Attitude
35	This apartment is on the first floor	**Okay** …. and I see it is close to the German church there	Pos	Imp	Attention
36	I don't know anything about such things	**Well okay**……………. yes but I think I'm happy with that	Neg	Exp	Attitude

10.4.3 Results

Positive or negative feedback was found in 18% of all user utterances in the AdApt database. Almost all subjects, 94%, used feedback at least once during their interaction with the system. User feedback occurred in a separate turn in as few as 6% of all cases in the AdApt corpus. Instead, feedback typically occurred in the initial position of a longer user sequence, after which a silent pause was followed by a request for more information. Turns 32 through 34 in the example dialogue in Table 10.5 provide examples of this phenomenon. In the AdApt database, 65% of all feedback utterances were judged to be positive. Two thirds of the feedback utterances were labeled as explicit, while one third were implicit. The groups of feedback tagged as attention or attitude were evenly sized.

When the function of the user feedback utterances was examined in a broader dialogue context, several interesting tendencies could be distinguished. The function of the largest group of utterances in the database was that of asking a direct question, for instance "finns det badkar" ('is there a bathtub') In these cases, feedback turned out to be quite uncommon. Another frequently occurring type of utterance in the database was one where the user would define his or her preferences. For this group, feedback was provided in about one fourth of the utterances. Relatively speaking, feedback was very frequent in those utterances that were used for concluding the interaction with the system. An example from the database is: "okej, då tackar jag för hjälpen" ('okay then, thanks for your help'). The feedback provided indicates that the user wants to sum up before finishing the dialogue. Meta–utterances, that is, user comments about the system, remarks on the preceding dialogue and self–directed communication, were quite rare in the corpus. When they occurred, however, they often included feedback to the system.

The analysis of data also revealed large individual variations in feedback strategies. While some subjects gave the system positive and negative feedback in virtually every turn, others very rarely gave feedback at all. For the individual subjects, the number of utterances that were labeled as including feedback varied from 0% to 70%. Figure 10.9 shows that about one fourth of the subjects used feedback in half or more of their turns, while one fourth of the subjects very rarely or never used feedback. No correlations with the subjects' reported experience with computers in general or spoken dialogue systems in particular were found. It appears as if feedback to a spoken dialogue system, at least partly, is a matter of individual style.

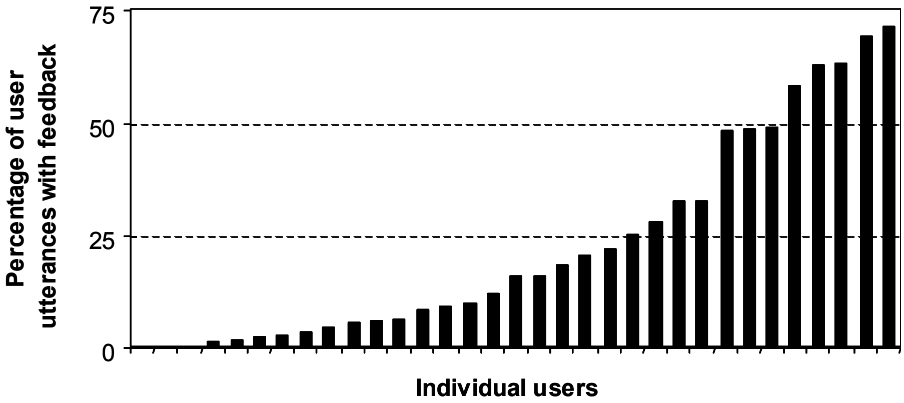


Figure 10.9 Distribution of feedback in user utterances.

10.4.4 Discussion and Summary

The human–computer dialogue as a whole probably affected the way in which feedback was used in the AdApt system. To investigate feedback in the context of the discourse, the system's previous turn was correlated to the users' choice of strategy. As can be seen in Figure 10.10, the feedback categories attention and attitude appeared at different places in the dialogue. In the initial phase of the discourse, where the system took the initiative and inquired about the user's preferences, feedback was often used to signal attitude. When the system failed to fulfill the user's request, on the contrary, users merely signalled that they had understood what the system was saying. Figure 10.10 also indicates that when the system turned over the initiative by asking an open question (e.g. 'Is there anything else you would like to know about the apartment?'), the subjects responded with attitude feedback: ('**Yes**, I would like to know if the apartment has a balcony'). It thus seems as if certain types of user feedback are likely to be provided in different phases in the dialogue. If, in a future system, it becomes apparent that difficulties often appear at a particular stage in the discourse, the system should anticipate negative user feedback. In this way, the user's warning to the system could perhaps prevent a more serious problem from occurring.

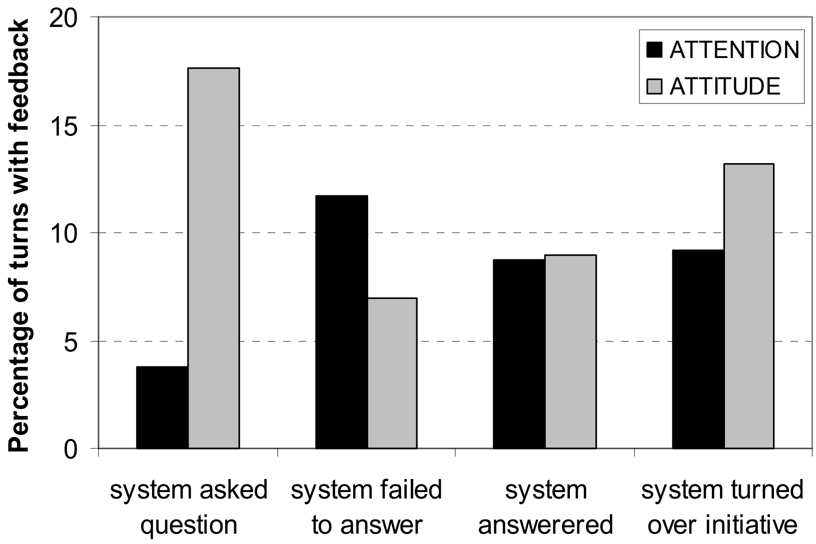


Figure 10.10 Number of turns with feedback depending on the previous system turn.

Those attitude feedback utterances that occurred after the system had supplied the user with information about some feature of an apartment, could be used to gain knowledge about the users' preferences. Instead of explicitly asking what kind of apartment the user would prefer, the system could attempt to interpret the user's feedback. For example, when a user asks: "Vad kan du berätta om den här lägenheten?" ('What can you tell me about this apartment?'), the system could present the apartment's most distinguishing feature(s). If the user provides the system with feedback, this could be used to decide which apartments to present later on in the dialogue. In Table 10.6, four examples of attitude feedback are presented. In two of the examples, the feedback might be used to model user preferences. In general, negative attitude feedback appeared to contain more information and be more useful than positive attitude feedback. For instance, the feedback utterance in the last example in the table could be used to detect that a problem has occurred in the dialogue, and that the user wishes to correct the system's interpretation.

Table 10.6 Translated examples of attitude feedback, marked for usability from the point of view of user preferences.

System answer to previous question	**User feedback**	**+/-**	**Usable**
The yellow building was constructed in 1890	**Yeah, that sounds good** – does it have an tiled stove maybe	Pos	Yes
The apartment is on the second floor	**Okay** and are there any available two-room apartments on Östermalm.	Pos	No
This building was constructed in 1997	**Ouch!** Is there an old building from the 19th century	Neg	Yes
This apartment has a tiled bathroom	**Well that's not quite what I asked about Urban** does it have a bathtub	Neg	No

In the present study, positive and negative user feedback cues were found to signal understanding and misunderstanding throughout the dialogues. We also found that certain user preferences were expressed in the feedback utterances. In a future system, positive feedback can perhaps be utilized as a way for the system to increase its knowledge about the user's preferences. Complicated correction subdialogues can thus be avoided. Negative feedback is sometimes used to warn the system of an upcoming problem. If these cues are correctly interpreted and handled by the system, serious errors can perhaps be prevented from occurring.

Part III

Chapter 11

Concluding Remarks

In this final chapter, we summarize the results from the empirical studies reported in Part II, and provide some examples of linguistic adaptation at each level. We then discuss some implications of these findings and suggest possible directions for future work.

11.1 Summary of Empirical Studies

At the phonetic level, we observed that speakers adapt their speech by modifying different acoustic-phonetic features. When compared to original utterances, verbatim repetitions in the August database were increased in duration, hyperarticulated and more often contained inserted pauses between words. There was also a tendency for children to increase the loudness of their speech during repetition. Similar tendencies were found in the Pixie database, where half of the adults' repetitions and almost two-thirds of the children's were labeled as hyperarticulated. An experiment with the purpose of investigating prosodic adaptation showed that subjects who interacted with the animated agent Cloddy Hans modified their speech rate towards that of the system's spoken output.

Lexical and syntactic adaptations in the August corpus were examined by comparing the database to another human–computer corpus, as well as a Swedish text database. We found few examples of syntactically complex structures in the data, something which can partly be explained by the fact that the average utterance in the August database was only four words. A study of near-repetitions in the database showed that speakers who rephrased their previous input often exchanged a lexical item for another, or tried different syntactically contrastive pairs. In the Cloddy Hans corpus, we investigated possible lexical entrainment effects. When the users' first utterances to the system were compared to their final ones, it was clear that the subjects had adapted their lexical choices to those of the system's output.

At the pragmatic level, studies of the August corpus showed that many users who interacted with this dialogue system were interested in other topics of conversation than information-seeking. More than 40% of all utterances in the corpus were categorized as socializing. A study of turn-taking in the AdApt corpus showed that many turns were fragmented, with inserted silent pauses. In a comparison between a spoken and a multimodal dialogue corpus, we noted that certain dialogue contexts appeared to more often elicit disfluencies in spontaneous language. In the multimodal AdApt corpus, these situations in the dialogue often occurred after the system had presented the user with a lot of information, and handed over the initiative. Finally, we observed that many users of the AdApt system were inclined to provide the system with positive and negative feedback, partly as a way of keeping the floor during certain parts of the discourse. For disfluencies as well as feedback, the individual differences were substantial.

As an illustration of the results presented in Part II, Table 11.1 exemplifies how speaker adaptation was realized at different linguistic levels in the four spoken dialogue corpora:

Table 11.1 Manifestations of linguistic adaptation in the corpora, at different levels.

Level	**Example of adaptation**
Phonetic/ Prosodic	Increase in loudness: *"Can you hear me AUGUST?"* Hyperarticulation: *"Where /pause/ does /pause/ the king /pause/ live?"*
Syntactic	Syntactically complex utterances during error handling: "*No now I misunderstood which button should I press the middle one or the one to the right*" Simplified language: *"Blue ball to red square"*
Lexical	Lexical convergence: *"The **yellow** one has a balcony" – "How much is the **yellow** one?"* Modification of lexical content: *"Do you have a cat?" – Do you have a kitty?"*
Pragmatic	Socializing: "*Does it make you sad that you don't understand?"* Feedback behavior: *"Great Urban /pause/ do you have another apartment in the area?"*

11.2 Discussion and Future Work

When we consider those of the spoken dialogue corpora that included children as well as adult speakers, a tendency can be seen. Adult users often try different strategies in their attempts to communicate with the system. For instance, they may systematically modify the lexical content and syntactic structures of their utterances. However, if the human–computer interaction becomes problematic, adult users often fall back on strategies acquired in human discourse, for instance by hyperarticulating. In general, adult behavioral strategies manifest a greater diversity and flexibility, and include many types of linguistic adaptations. Children's behavior, on the other hand, tends to be less varied. For example, in our corpora children users often repeat the same utterance over and over again while adapting phonetically and/or prosodically. Another tendency is that when young speakers have been misunderstood, they tend to stick to the same adaptive behavioral pattern rather than trying out different ways of making themselves understood.

To what extent are the linguistic adaptations examined in the four corpora unique strategies of human–computer interaction? Expert users of spoken dialogue systems may adapt their language in accordance with previously attempted strategies, especially if they turned out to be successful. However, most users who adapt their language do so without knowing much about the spoken dialogue system's capabilities, or the possible effects of various types of speech input. In these cases, it can be argued that adapting to a computer is no different than adapting to another human being who appears to have problems understanding. Just as when addressing an elderly person or a non-native speaker, the user tries to conceptualize what sort of language input the dialogue system is able to handle and respond to.

If we are to achieve the goal of creating a computer system with conversational abilities, it is necessary to understand and be able to model users' linguistic adaptations. Knowledge in the area is scattered, and its incorporation in speech technology systems has to date been limited. At the same time, the relative immaturity of current spoken dialogue systems in itself frequently causes errors and miscommunication in human–computer dialogue. Linguistic adaptations at different levels are often indicators of errors in dialogue, and if quickly identified they can be used to prevent serious problems from occurring. This motivates studies of user adaptation with the dual purpose of contributing to general knowledge about users' behavior but also to help incrementally improve spoken dialogue systems.

Some parts of the adaptive process, such as the effect of the system's output on the users' lexical and syntactical choices, have been relatively well understood. However, future studies will be needed to investigate how other aspects of the interface design and dialogue context contribute to user adaptations. Explicitly instructing users on how to speak to an experimental dialogue system should be avoided, especially considering the long-term goal of creating interfaces that can handle unrestricted natural language input. Instead, designers of future systems will have to take focus on how speakers naturally behave, and to some extent on how they can be subtly influenced to adapt their language to match that of current systems' capacities. It is clear that at the lexical, syntactic and to some extent phonetic/prosodic levels, it is possible to achieve such convergence effects.

Spoken dialogue systems with certain features will elicit social user behavior. Although some knowledge has been gained in this area, a lot of research still lies ahead of us in the area of modelling users' linguistic adaptations at the pragmatic level. As we develop conversational spoken dialogue systems which integrate an increasing number of human discourse features, this will require more advanced turn-handling capabilities, sophisticated social skills and even the ability to respond to irony.

Users of spoken dialogue systems are not all alike. Individual differences in speaker behavior are significant, and will probably become more accentuated as we develop conversational systems which allow for greater degrees of user initiative. Users favor different ways of speaking, and follow different linguistic adaptive patterns during problematic dialogues and error resolution. Ideally, a spoken dialogue system should be able to handle these personal variations, something which should make the interaction more successful. An area in which much future work remains to be done is in modelling individual differences in linguistic adaptations.

List of Publications

Bell, L., Gustafson, J. and Heldner, M. (2003). Prosodic adaptation in human–computer interaction. *Proceedings of ICPhS'03*, 2453-2456.

Bell, L. and Gustafson, J. (2003). Child and adult speaker adaptation during error resolution in a publicly available spoken dialogue system. *Proceedings of Eurospeech '03*, 613-616.

Gustafson, J., Bell, L., Boye, J., Edlund, J. & Wirén, M. (2002). Constraint manipulation and visualization in a multimodal dialogue system. *Proceedings of the ISCA Workshop on Multimodal Dialogue in Mobile Environments* Kloster Irsee, Germany.

Bell, L., Boye, J. & Gustafson, J. (2001). Real-time handling of fragmented utterances. *Proceedings of the NAACL 2001 Workshop on Adaptation in Dialogue Systems*, Pittsburgh, USA.

Bell, L. (2000). *Linguistic Adaptations in Spoken and Multimodal Dialogue Systems*. Licentiate thesis, Centre for Speech Technology, KTH, December 2000.

Bell, L., Boye, J., Gustafson, J. & Wirén, M. (2000). Modality convergence in a multimodal dialogue system. *Proceedings of Götalog 2000*, Fourth Workshop on the Semantics and Pragmatics of Dialogue, 29–34.

Gustafson, J. & Bell, L. (2000). Speech technology on trial: Experiences from the August System. *Natural Language Engineering*: Special Issue on Best Practice in Spoken Dialogue Systems, 273–286.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. & Wirén M. (2000). AdApt – a multimodal conversational dialogue system in an apartment domain." *Proceedings of ICSLP '00*, 134–137.

Bell, L. & Gustafson, J. (2000). Positive and negative user feedback in a spoken dialogue corpus. *Proceedings of ICSLP 00*, 589–592.

Bell, L., Eklund, R. & Gustafson, J. (2000). A comparison of disfluency distribution in a unimodal and a multimodal speech interface. *Proceedings of ICSLP '00*, 626–629.

Bell, L. and Gustafson, J. (1999a). Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer directed speech. *Proceedings of ICPhS '99*, 1221–1224.

Bell, L. and Gustafson, J. (1999b). Interaction with an animated agent in a spoken dialogue system. *Proceedings of Eurospeech* '99, 1143–1146.

Bell, L. & Gustafson, J. (1999c). Utterance types in the August System. *Proceedings of the ESCA Workshop on Interactive Dialogue in Multimodal Systems*: 81-84.

References

Ainsworth, W. A. and Pratt, S. R. (1992). Feedback strategies for error correction in speech recognition systems. *International Journal of Man-Machine Studies* **36**: 833-842.

Allen, J., Ferguson, G. and Stent, A. (2001). An architecture for more realistic conversational systems. In *Proceedings of Intelligent User Interfaces***:** 1-8.

Allen, J. F., Miller, B. W., Ringger, E. K. and Sikorski, T. (1996). Robust understanding in a dialogue system. In *Proceeding of 34th meeting of the Association for Computational Linguistics*.

Alleva, F., Huang, X., Hwang, M.-Y. and Jiang, L. (1997). Can continuous speech recognizers handle isolated speech? In *Proceedings of Eurospeech'97*.

Allwood, J. (1996). Några perspektiv på talspråksforskning. In *Samspel och variation: Språkliga studier tillägnade Bengt Nordberg på 60-årsdagen*, Department of Nordic Languages, Uppsala University.

Allwood, J. (2000). An activity based approach to pragmatics. In *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. H. Bunt and B. Black Eds. Amsterdam, John Benjamins**:** 47-80.

Allwood, J. and Haglund, B. (1992). Communicative activity analysis of a Wizard of Oz experiment. In *Internal report. PLUS ESPRIT project P5254*.

Allwood, J., Nivre, J. and Ahlsén, E. (1993). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics* **9**(1).

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech* **34**: 351-366.

Arunachalam, S., Gould, D., Andersen, E., Byrd, D. and Narayanan., S. S. (2001). Politeness and frustration language in child-machine interactions. In *Proceedings of Eurospeech***:** 2675-2678.

Aust, H., Oerder, M., Seide, F. and Steinbiss, V. (1995). The Philips automatic train timetable information system. *Speech Communication* **17**(3-4): 249-262.

Austin, J. L. (1961). Performative utterances. In *The Philosophy of Language, 3rd ed.* A. P. Martinich Ed. New York, Oxford University Press**:** 120-129.

Balentine, B. and Morgan, D. P. (1999). *How to Build a Speech Recognition Application: A Style Guide for Telephony Dialogues*. San Ramon, Ca, Enterprise Integration group.

Beattie, G. (1983). *Talk: An Analysis of Speech and Non-verbal Behaviour in Conversation*, Open University Press.

Bell, L., Boye, J. and Gustafson, J. (2001). Real-time handling of fragmented utterances. In *Proceedings of the NAACL 2001 Workshop on Adaptation in Dialogue Systems* Pittsburgh, PA.

Bell, L., Boye, J., Gustafson, J. and Wirén, M. (2000). Modality convergence in a multimodal dialogue system. In *Proceedings of Götalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue*: 29-34.

Bell, L. and Gustafson, J. (1999a). Interaction with an animated agent in a spoken dialogue system. In *Proceedings of Eurospeech '99* Budapest. **3:** 1143-1146.

Bell, L. and Gustafson, J. (1999b). Utterance types in the August System. In *Proceedings of the ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*: 81-84.

Bell, L. and Gustafson, J. (2000). Positive and negative user feedback in a spoken dialogue corpus. In *Proceedings of ICSLP '00* Beijing.

Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., Serpa-Leitao, A. and Ström, N. (1995a). The Waxholm system - a progress report. In *Proceedings of ESCA Tutorial and Workshop on Spoken Dialogue Systems* Vigsö, Denmark**:** 81-84.

Bertenstam, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., Serpa-Leitao, A. and Ström, N. (1995b). Spoken dialogue data collected in the Waxholm project. *STL-QPSR, Department of Speech Communication* **1**: 50-73.

Beskow, J. (2003). *Talking Heads - Models and Applications for Multimodal Speech Synthesis*. Stockholm, PhD Thesis, KTH.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge, Cambridge University Press.

Bickmore, T. and Cassell, J. (2001). Relational agents: a model and implementation of building user trust. In *Proceedings of CHI 2001*.

Bickmore, T. and Cassell, J. (forthcoming). Social dialogue with embodied conversational agents. In *Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*. J. v. Kuppevelt, L. Dybkjaer and N. O. Bernsen Eds.

Bloomfield, L. (1933). *Language*. New York, Holt, Rinehart and Winston.

Blum-Kulka, S. (1997). Discourse pragmatics. In *Discourse as Social Interaction*. T. Van Dijk Ed. London, Sage Publications.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F. and Brennan, S. E. (1999). Which speakers are most disfluent, and when? In *Proceedings of ICPhS'99 Satellite Meeting on Disfluency in Spontaneous Speech* U C Berkeley.

Boulard, H., Hermansky, H. and Morgan, N. (1996). Towards increasing speech recognition error rates. *Speech Communication* **18**: 205-208.

Boyce, S. J. (1999). Spoken natural language dialogue systems: user interface issues for the future. In *Human Factors and Voice Interactive Systems*. D. Gardner-Bonneau Ed. Norwell, Massachusetts, Kluwer Academic Publishers**:** 37-62.

Boye, J., Hockey, B. A. and Rayner, M. (2000). Asynchronous dialogue management: Two case-studies. In *Proceedings of Götalog. Fourth Workshop on the Semantics and Pragmatics of Dialogue***:** 51-55.

Branigan, H., Lickley, R. and McKelvie, D. (1999). Non-linguistic influences on rates of disfluency in spontaneous speech. In *Proceedings of ICPhS'99* San Francisco**:** 387-390.

Branigan, H. P., Pickering, M. J. and Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition* **75**: B13-25.

Bredvad-Jensen, A.-C. (1995). Prosodic variation in parental speech in Swedish. In *Proceedings of ICPhS* Stockholm.

Brennan, S. (1990). Conversation as direct manipulation: an iconoclastic view. In *The Art of Human–Computer Interface Design*. B. Laurel Ed., Addison-Wesley**:** 393-404.

Brennan, S. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of ISSD***:** 41-44.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning. Memory and Cognition* **22**(6): 1482-1493.

Brennan, S. E. and Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: a theoretical framework. *Knowledge-Based Systems*(8): 143-151.

Brown, P. and Levinson, S. (1987). *Politeness*, Cambridge University Press.

Bull, M. (1996). An analysis of between-speaker intervals. In *Proceedings of the Edinburgh Linguistic Department Conference***:** 18-27.

Carlberger, A. (1999). Nparse - a shallow n-gram-based grammatical-phrase parser. *STL-QPSR, Department of Speech, Music and Hearing* **38**(3-4): 1-10.

Carlson, R., Granström, B. and Hunnicutt, S. (1982). A multi-language text-to-speech module. In *Proc. ICASSP-Paris*. **3:** 1604-1607.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of CHI*: 520-527.

Cettolo, M. and Falavigna, D. (1998). Automatic detection of semantic boundaries based on acoustic and lexical knowledge. In *Proceedings of ICSLP*: 1551-1554.

Chapanis, A. (1981). Interactive human communication: Some lessons learned from laboratory experiments. In *Man-Computer Interaction: Human Factors Aspects of Computers and People*. B. Shackel Ed. Rockville, Maryland, Sijthoff & Noordhoff.

Cheepen, C. (1988). *The Predictability of Informal Conversation*. London, New York, Pinter Publishers.

Cheepen, C. (1996). Guidelines for dialogue design - what is our approach? *Working paper No 3, ESRC-funded project Guidelines for Advanced Voice Dialogues*.

Clark, H. H. (1994). Managing problems in speaking. *Speech Communication* **15**(3-4): 243-250.

Clark, H. H. (1996). *Using Language*. Cambridge, Cambridge University Press.

Clark, H. H. and Schaefer, E. W. (1989). Contributing to discourse. *Cognitive Science* **13**: 259-294.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* **22**: 1-39.

Core, M. G. and Allen, J. F. (1997). Coding dialogs with the DAMSL annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines* Boston, MA.

Coulston, R., Oviatt, S. and Darves, C. (2002). Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of ICSLP'02* Denver, CO.

Dahlbäck, N. (1991). *Representations of Discourse - Cognitive and Computational Aspects*, PhD Thesis, Linköping University.

Darves, C. and Oviatt, S. (2002). Adaptation of users' spoken dialogue patterns in a conversational interface. In *Proceedings of ICSLP'02* Denver, CO.

Day, M. C. and Boyce, S. J. (1993). Human factors in human–computer system design. In *Advances in Computers*. M. C. Yovits Ed. San Diego, Academic Press. **36:** 333-430.

Drew, P. and Sorjonen, M.-L. (1997). Institutional dialogue. In *Discourse as Social Interaction*. T. Van Dijk Ed. London, Sage Publications.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and Vreken, O. v. d. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In *Proceedings of ICSLP'96* Philadelphia. **3:** 1393-1396.

Dybkjaer, H., Bernsen, N. O. and Dybkjaer, L. (1993). Wizard-of-Oz and the trade-off between naturalness and recogniser constraints. In *Proceedings of Eurospeech***:** 947-950.

Edgington, M., Attwater, D. and Durston, P. (1999). Oasis - a framework for spoken language call steering. In *Proceedings of Eurospeech*.

Edlund, J. and Nordstrand, M. (2002). Turn-taking gestures and hourglasses in a multi-modal dialogue system. In *Proceedings of the ISCA Workshop on Multi-Modal Dialogue in Mobile Environments* Kloster Irsee, Germany.

Egan, D. E. (1988). Individual differences in human–computer interaction. In *Handbook of Human–Computer Interaction*. M. Helander Ed. Amsterdam, North-Holland.

Eklund, R. (1999). A comparative study of disfluencies in four Swedish travel dialogue corpora. In *Proceedings of Disfluency in Spontaneous Speech Workshop* Berkeley, California**:** 3-6.

Eskenazi, M., Rudnicky, A., Gregory, K., Constantinides, P., Brennan, R., Bennett, C. and Allen, J. (1999). Data Collection and Processing in the Carnegie Mellon Communicator. In *Proceedings of Eurospeech '99***:** 2695-2698.

Fillmore, C., Ed. (1981). *Pragmatics and the description of discourse*. Radical Pragmatics. New York, Academic Press.

Fraser, N. and Gilbert, N. (1991). Simulating speech systems. *Computer Speech and Language* **5**: 81-99.

Fromkin, V. A., Ed. (1980). *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand*. New York, Academic Press.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM* **30**(11): 964-971.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* **27**: 181-218.

Giles, H., Mulac, A., Bradac, J. and Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. In *Communication Yearbook 10*. M. L. McLaughlin Ed.

Godfrey, J. J., Holliman, E. C. and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE* San Francisco**:** 517-520.

Gorin, A., Riccardi, G. and Wright, J. (1997). How May I Help You? *Speech Communication* **23**: 113-127.

Gould, J. D., Conti, J. and Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM* **26**(4): 295-308.

Greenberg, S. (1998). Recognition in a new key - Towards a science of spoken language. In *Proceedings of ICASSP***:** 1041-1045.

Grice, H. P. (1975). Logic and Conversation. In *The Philosophy of Language*. A. P. Martinich Ed. New York, Oxford University Press**:** 156-167.

Guindon, R., Shuldberg, K. and Conner, J. (1987). Grammatical and ungrammatical structures in user-advised dialogues: Evidence for sufficiency of restricted languages in natural language interfaces to advisory systems. In *Proceedings of ACL***:** 41-44.

Gustafson, J. (2002). *Developing Multimodal Spoken Dialogue Systems - Empirical Studies of Spoken Human–Computer Interaction*. Stockholm, PhD Thesis, KTH.

Gustafson, J. and Bell, L. (2000). Speech technology on trial - Experiences from the August system. *Natural Language Engineering* **6**(3-4): 273-286.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. and Wirén, M. (2000). AdApt – a multimodal conversational dialogue system in an apartment domain. In *Proceedings of ICSLP '00*. **II:** 134-137.

Gustafson, J., Larsson, A., Carlson, R. and Hellman, K. (1997). How do system questions influence lexical choices in user answers? In *Proceedings of Eurospeech*.

Gustafson, J., Lindberg, N. and Lundeberg, M. (1999). The August spoken dialogue system. In *Proceedings of Eurospeech '99***:** 1151-1154.

Gustafson, J., Lundeberg, M. and Liljencrants, J. (1999). Experiences from the development of August - a multimodal spoken dialogue system. In *Proceedings of IDS*.

Gustafson, J. and Sjölander, K. (2002). Voice Transformations for Improving Children's Speech Recognition in a Publically Available Dialogue System. In *Proceedings of ICSLP'02*. **1:** 297-300.

Halliday, M. A. K. (1989). *Spoken and Written Language*, Oxford University Press.

Hauptmann, A. and Rudnicky, A. (1990). A comparison of speech and typed input. In *Proceedings of the Speech and Natural Language Workshop* San Mateo, Ca, Morgan Kaufmann Publishers, Inc.**:** 219-224.

Heeman, P. (1997). *Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog*, PhD Thesis, University of Rochester.

Heeman, P. and Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics* **21**(3).

Heeman, P. A. and Allen, J. F. (1999). Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics* **25**(4).

Helander, M., Landauer, T. K. and Prabhu, P., Eds. (1997). *Handbook of Human–Computer Interaction*, Elsevier Science.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics***:** 123–128.

Hirschberg, J., Litman, D. and Swerts, M. (1999). Prosodic cues to recognition errors. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop***:** 349-352.

Hirschberg, J. and Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of ACL***:** 286-293.

Hirschman, L., Bates, M., Dahl, D., Fisher, W., Garofolo, J., Pallett, D., Hunike-Smith, K., Price, P., Rudnicky, A. and Tzoukermann, E. (1992). Multi-site data collection and evaluation in spoken language understanding. P. o. t. D. S. a. N. L. Workshop Ed. Harriman, NY, Morgan Kaufmann Publishers, Inc.**:** 7-14.

Hirst, G., McRoy, S., Heeman, P., Edmonds, P. and Horton, D. (1994). Repairing conversational misunderstandings and non-understandings. *Speech Communication*(15): 213-229.

Hotopf, W. H. N. (1980). Semantic similarity as a factor in whole-word slips of the tongue. In *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand.* V. A. Fromkin Ed. New York, Academic Press**:** 97-109.

HSFR (1994). Etik - forskningsetiska principer för humaniora och samhällsvetenskap.

Hunnicutt, S., Hirschman, L., Polifroni, J. and Seneff, S. (1992). Analysis of the Effectiveness of System Error Messages in a Human-Machine Travel Planning Task. In *Proceedings of ICSLP '92*. **1:** 197-200.

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing*, Prentice Hall.

Jönsson, A. (1993). *Dialogue Management for Natural Language Interfaces*, PhD Thesis. Linköping Studies in Science and Technology 312.

Jönsson, A. and Dahlbäck, N. (1988). Talking to a computer is not like talking to your best friend. In *Proceedings of the first Scandinavian conference on artificial intelligence* Tromsö, Norway.

Kamm, C. and Helander, M. (1997). Design issues for interfaces using voice input. In *Handbook of Human–Computer Interaction*. M. G. Helander, T. K. Landauer and P. V. Prabhu Eds. Amsterdam, Elsevier Science.

Karat, J., Lai, J., Danis, C. and Wolf, C. (1999). Speech user interface evolution. In *Human Factors and Voice Interactive Systems*. D. Gardner-Bonneau Ed. Norwell, Massachusetts, Kluwer Academic Publishers**:** 1-35.

Karl, L., Pettey, M. and Shneiderman, B. (1993). Speech-activated versus mouse-activated commands for word processing applications: An empirical study. *International Journal of Man-Machine Studies* **39**(4): 667-687.

Karlsson, F. (1992). SWETWOL: a comprehensive morphological analyser for Swedish. *Nordic Journal of Linguistics* **15**: 1-45.

Kennedy, A., Wilkes, A., Elder, L. and Murray, W. (1988). Dialogue with machines. *Cognition* **30**: 73-105.

Krahmer, E., Swerts, M., Theune, M. and Weegels, M. (1999). Problem spotting in human-machine interaction. In *Proceedings of Eurospeech '99*. **3:** 1423-1426.

Krahmer, E., Swerts, M., Theune, M. and Weegels, M. (2001). Error detection in spoken human-machine interaction. *International Journal of Speech Technology* **4**(1): 19-30.

Krahmer, E., Swerts, M., Theune, M. and Weegels, M. (2002). The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Communication* **36**(1): 133-145.

Lamel, L., Bennacef, S., Gauvain, J. L., Dartigues, H. and Temem, J. N. (1998). User evaluation of the mask kiosk. In *Proceedings of ICSLP '98***:** 2875-2878.

Lamel, L. F., Bennacef, S. K., Rosset, S., Devillers, L., Foukia, S., Gangolf, J. J. and Gauvain, J. L. (1997). The LIMSI RailTel system: field trial of a telephone service for rail travel information. *Speech Communication* **23**(1-2): 67-82.

Landauer, T. K. (1997). Behavioral research methods in human–computer interaction. In *Handbook of Human–Computer Interaction*. P. Prabhu Ed., Elsevier Science.

Langkilde, I., Walker, M., Wright, J., Gorin, A. and Litman, D. (1999). Automatic prediction of problematic human–computer dialogues in 'How may I help you?' In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*.

Lee, S., Potamianos, A. and Narayanan, S. (1999). Acoustics of children's speech: developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America* **105**: 1455-1468.

Leech, G. (1991). The state of the art in corpus linguistics. In *English Corpus Linguistics*. K. Aijmer and B. Altenberg Eds. London and New York, Longman.

Lester, J., Converse, S., Kahler, A., Barlow, T., Stone, B. and Bhoga, R. (1997). The persona effect: affective impact of animated pedagogical agents. In *Proceedings of CHI*.

Levelt, W. J. M. and Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology* **14**: 78-106.

Levinson, S. C. (1983). *Pragmatics*, Cambridge University Press.

Levinson, S. E. (1994). Speech recognition technology: a critique. In *Voice Communication Between Humans and Machines*. D. B. Roe and J. G. Wilpon Eds., National Academy of Sciences**:** 159-164.

Levow, G.-A. (1998). Characterizing and recognizing spoken corrections in human–computer dialogue. In *Proceedings of COLING-ACL '98*.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech* **6**: 172-187.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In *Speech Production and Speech Modelling*. W. J. Hardcastle and A. Marchal Eds. Dordrecht, Kluwer Academic Publishers**:** 403-439.

Lundeberg, M. and Beskow, J. (1999). Developing a 3D-agent for the August dialogue system. In *Proceedings of ESCA Workshop on Audio-Visual Speech Processing***:** 151-154.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* **63**: 81-97.

Miller, J. and Weinert, R. (1998). *Spontaneous Spoken Language - Syntax and Discourse*. Oxford, Oxford University Press.

Moon, S.-J. and Lindblom, B. (1989). Formant undershoot in clear and citation-form speech: a second progress report. *STL-QPSR, Department of Speech Communication* **1**.

Nakano, M., Miyazaki, N., Hirasawa, J., Dohsaka, K. and Kawabata, T. (1999). Understanding unsegmented user utterances in real-time spoken dialogue systems. In *Proceedings of ACL***:** 200-207.

Nakatani, C. and Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* **95**(3): 1603-1616.

Narayanan, S. and Potamianos, A. (2002). Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing* **10**(2): 65-78.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *Journal of Social Issues* **56**(1): 81-103.

Nass, C. and Steuer, S. (1993). Voices, boxes, and sources of messages: computers and social actors. *Human Communications Research* **19**: 504-527.

Nord, L. (1986). Acoustic studies of vowel reduction in Swedish. *KTH-QPSR* **4**: 19-36.

Nuance Application Developer's Guide, *v.* (2000).

Oviatt, S. (1995). Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Language* **9**: 19-35.

Oviatt, S. and Adams, B. (2000). Designing and evaluating conversational interfaces with animated characters. In *Embodied Conversational Agents*. J. Cassell, J. Sullivan, S. Prevost and E. Churchill Eds. Cambridge, MA, MIT Press**:** 319-343.

Oviatt, S., Bernard, J. and Levow, G.-A. (1998a). Linguistic adaptations during spoken and multimodal error resolution. *Language and Speech* **41**(3-4): 419-442.

Oviatt, S., Levow, G.-A., MacEachern, M. and Kuhn, K. (1996). Modeling hyperarticulate speech during human–computer error resolution. In *Proceedings of ICSLP*. **2:** 801-804.

Oviatt, S., Levow, G.-A., Moreton, E. and MacEachern, M. (1998b). Modeling global and focal hyperarticulation during human–computer error resolution. *The Journal of the Acoustical Society of America* **104**(5): 3080-3098.

Oviatt, S., MacEachern, M. and Levow, G.-A. (1998c). Predicting hyperarticulate speech during human–computer error resolution. *Speech Communication* **24**(2): 1-23.

Oviatt, S. and VanGent, R. (1996). Error resolution during multimodal human–computer interaction. In *Proceedings of ICSLP'96*: 204-207.

Picheny, M. A., Durlach, N. I. and Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research* **28**: 96-103.

Picheny, M. A., Durlach, N. I. and Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research* **29**: 434-446.

Pierce, J. R. (1969). Whither speech recognition? *JASA* **46**(4): 1049--1051.

Potamianos, A., Narayanan, S. and Lee, S. (1997). Automatic speech recognition for children. In *Proceedings of Eurospeech*. **5:** 2371-2374.

Reeves, B. and Nass, C. (1996). *The Media Equation: How people treat computers, television and new media like real people and places*. New York, Cambridge University Press.

Reilly, R. G., Ed. (1987). *Communication Failure in Dialogue and Discourse*, Elsevier Science.

Rhyne, J. R. and Wolf, C. G. (1993). Recognition-based user interfaces. In *Advances in Human–Computer Interaction*. H. R. Hartson and D. Hix Eds. Norwood, N.J., Ablex Publishing. **4:** 191-250.

Sacks, H., Schegloff, E. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* **50**(4): 696-735.

Schegloff, E. A. and Sacks, H. (1973). Opening up closings. *Semiotica* **7**(4): 289-327.

Searle, J. R. (1965). What is a speech act? In *The Philosophy of Language, 3rd ed.* A. P. Martinich Ed. New York, Oxford University Press**:** 130-140.

Searle, J. R. (1979). A taxonomy of illocutionary acts. In *The Philosophy of Language, 3rd ed.* A. P. Martinich Ed. New York, Oxford University Press**:** 141-155.

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM* **43**(9): 63-65.

Shriberg, E. (1996). Disfluencies in Switchboard. In *Proceedings of ICSLP***:** 11-14.

Shriberg, E., Wade, E. and Price, P. (1992). Human-machine problem solving using spoken language systems: Factors affecting performance and user satisfaction. In *Proceedings of the DARPA Speech and Natural Language Workshop* San Mateo, CA, Morgan Kaufmann Publishers**:** 49-54.

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*, PhD Thesis, University of California at Berkeley.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). TOBI: a standard for labeling English prosody. In *Proceedings of ICSLP*.

Sjölander, K. (2001). Automatic alignment of phonetic segments. In *Working papers 49: Proceedings of Fonetik 2001* Lund, Lund University, Department of Linguistics: 140-143.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. *Proceedings of Fonetik*: 93-96.

Skantze, G. (2002). Coordination of referring expressions in multimodal human–computer dialogue. In *Proceedings of ICSLP* Denver, CO.

Smith, R. E. (1993). *Psychology*, West publishing company.

Smith, R. W. (1998). An evaluation of strategies for selectively verifying utterance meanings in spoken natural language dialog. *International Journal of Human–Computer Studies* **48**: 627-647.

Smith, R. W. and Gordon, S. A. (1996). Pragmatic issues in handling miscommunication: observations of a spoken natural language dialog system. In *Proceedings of the AAAI '96 Workshop on Detecting, Repairing and Preventing Human-Machine Miscommunication*: 21-28.

Smith, R. W. and Hipp, D. R. (1994). *Spoken Natural Language Dialogue Systems: A Practical Approach*. New York, Oxford University Press.

Soltau, H. and Waibel, A. (2002). Acoustic models for hyperarticulated speech. In *Proceedings of ICSLP*.

Stalnaker, R. C. (1978). Assertion. In *Syntax and semantics, vol 9: Pragmatics*. P. Cole Ed. New York, Academic Press: 315-332.

Stent, A. J. (2001). *Dialogue systems as conversational partners: applying conversation acts theory to natural language generation for task-oriented mixed-initiative spoken dialogue*, PhD Thesis, University of Rochester.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykma, C. V., Ries, K., Shriberg, E., Jurafsky, D., Martin, R. and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* **26**(3).

Stolcke, A. and Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. *Proceedings of ICSLP*: 1005-1008.

Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tur, G. and Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of ICSLP***:** 2247-2250.

Ström, N. (1997). *Automatic Continuous Speech Recognition with Rapid Speaker Adaptation for Human/Machine Interaction.* Stockholm, PhD Thesis, Department of Speech, Music and Hearing, KTH.

Ström, N. and Seneff, S. (2000). Intelligent barge-in in conversational systems. In *Proceedings of ICSLP.*

Suhm, B. and Waibel, A. (1997). Exploiting repair context in interactive error recovery. In *Proceedings of Eurospeech '97.*

Sundberg, U. (1998). *Mother Tongue – Phonetic Aspects of Infant-Directed Speech*, PhD Thesis, Department of Linguistics, Stockholm University.

Svartvik, J. and Quirk, R. (1980). *A Corpus of English Conversation.* Lund, Gleerup.

Swerts, M., Koiso, H., Shimojima, A. and Katagiri, Y. (1998). On different functions of repetitive utterances. In *Proceedings of ICSLP* Sydney. **4:** 1287-1290.

Thomas, J. C., Ed. (1995). *Human factors in lifecycle development.* Applied Speech Technology. Boca Raton, CRC Press.

Traum, D. (1994). *A Computational Theory of Grounding in Natural Language Conversation*, PhD Thesis, University of Rochester.

Traum, D. and Heeman, P. (1997). Utterance units in spoken dialogue. In *Processing in Spoken Dialogue Systems.* E. Maier, M. Mast and S. LuperFoy Eds.**:** 125-140.

Traum, D. R. and Dillenbourg, P. (1996). Miscommunication in multi-modal collaboration. In *Proceedings of the AAAI '96 Workshop on Detecting, Reparing and Preventing Human-Machine Miscommunication.*

van Bergem, D. (1993). Acoustic vowel reduction as a function of sentence accent, word stress and word class. *Speech Communication* **12**: 1-23.

van den Broecke, M. P. R. and Goldstein, L. (1980). Consonant features in speech errors. In *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand.* V. A. Fromkin Ed. New York, Academic Press**:** 47-65.

Ward, K. and Heeman, P. (2000). Acknowledgements in human–computer interaction. In *Proceedings of ANLP-NAACL***:** 281-287.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* **91**(3): 1707-1717.

Wilpon, J. and Jacobsen, C. (1996). A study of speech recognition for children and the elderly. In *Proceedings of ICASSP* Atlanta, GA, IEEE Press**:** 349-352.

Wooffitt, R. and MacDermid, C. (1995). Wizards and social control. In *The Social and Interactional Dimensions of Human–Computer Interaction*. P. J. Thomas Ed. Cambridge, Cambridge University Press.

Yankelovich, N. (1996). How Do Users Know What to Say? *ACM Interactions* **3**(6): 33-43.

Yankelovitch, N., Levow, G. and Marx, M. (1995). Designing speech acts: issues in speech user interfaces. In *Proceedings of the Conference on Human Factors in Computer Systems* Denver, CO, ACM Press**:** 369-376.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies* **34**: 527-547.

Zue, V. (2000). JUPITER: a telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing* **8**(1).