Developing Multimodal Spoken Dialogue Systems

Empirical Studies of Spoken Human–Computer Interaction

Joakim Gustafson

Doctoral Dissertation
Stockholm
2002

Developing Multimodal Spoken Dialogue Systems

Empirical Studies of Spoken Human–Computer Interaction

Joakim Gustafson

Doctoral Dissertation
Department of Speech, Music and Hearing, KTH
Stockholm 2002

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknisk doktorsexamen fredagen den 20 december 2002 kl 14.00 i Kollegiesalen, Administrationsbyggnaden, KTH, Valhallavägen 79, Stockholm.

TRITA-TMH 2002:8
ISSN 1104-5787

Cover picture, “A hard day’s night in the life of August”, by the author
Typeset by the author using Microsoft® Word
Printed by Universitetsservice US AB, Stockholm, 2002

Abstract

This thesis presents work done during the last ten years on developing five multimodal spoken dialogue systems, and the empirical user studies that have been conducted with them. The dialogue systems have been multimodal, giving information both verbally with animated talking characters and graphically on maps and in text tables. To be able to study a wider rage of user behaviour each new system has been in a new domain and with a new set of interactional abilities. The five system presented in this thesis are: The *Waxholm* system where users could ask about the boat traffic in the Stockholm archipelago; the *Gulan* system where people could retrieve information from the Yellow pages of Stockholm; the *August* system which was a publicly available system where people could get information about the author Strindberg, KTH and Stockholm; the *AdApt* system that allowed users to browse apartments for sale in Stockholm and the *Pixie* system where users could help an animated agent to fix things in a visionary apartment publicly available at the Telecom museum in Stockholm. Some of the dialogue systems have been used in controlled experiments in laboratory environments, while others have been placed in public environments where members of the general public have interacted with them. All spoken human-computer interactions have been transcribed and analyzed to increase our understanding of how people interact verbally with computers, and to obtain knowledge on how spoken dialogue systems can utilize the regularities found in these interactions. This thesis summarizes the experiences from building these five dialogue systems and presents some of the findings from the analyses of the collected dialogue corpora.

Keywords: Spoken dialogue system, multimodal, speech, GUI, animated agents, embodied conversational characters, talking heads, empirical user studies, speech corpora, system evaluation, system development, Wizard of Oz simulations, system architecture, linguistic analysis.

Table of Contents

Included papers

The second part of this dissertation consists of the following papers:

Paper I Bertenstam, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa–Leitao, A., Nord, L. and Ström, N.
Spoken dialogue data collection in the Waxholm project
STL-QPSR 1/1995, pp. 50–73, 1995.

Paper II Gustafson, J., Larsson, A., Carlson, R. and Hellman, K.
How do System Questions Influence Lexical Choices in User Answers?
Proceedings of Eurospeech 97, pp. 2275–2278, Rhodes, Greece, 1997.

Paper III Bell, L. and Gustafson, J.
Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech
Proceedings of ICPhS 99, vol. 2 pp. 1221–1224, San Francisco, USA, 1999.

Paper IV Gustafson, J. and Bell, L.
Speech Technology on Trial: Experiences from the August System
Journal of Natural Language Engineering: Special issue on Best Practice in Spoken Dialogue Systems, pp. 273–286, 2000.

Paper V Bell, L., Boye, J., Gustafson, J., and Wirén, M.
Modality Convergence in a Multimodal Dialogue System
Proceedings of Götalog, Fourth Workshop on the Semantics and Pragmatics of Dialogue, pp. 29–34, Göteborg, Sweden, 2000.

Paper VI Bell, L. and Gustafson, J.
Positive and Negative User Feedback in a Spoken Dialogue Corpus
Proceedings of ICSLP 00, vol. 1, pp. 589–592, Beijing, China, 2000.

Paper VII Bell, L., Eklund, R. and Gustafson, J.
A Comparison of Disfluency Distribution in a Unimodal and a Multimodal Speech Interface
Proceedings of ICSLP 00, vol. 3, pp. 626–629, Beijing, China, 2000.

Paper VIII Bell, L., Boye, J., and Gustafson, J.
Real-time Handling of Fragmented Utterances
Proceedings of the NAACL 01 workshop on Adaptation in Dialogue Systems, Pittsburgh, USA, 2001.

Paper IX Gustafson, J., Bell, L., Boye, J., Edlund, J. and Wirén, M.
Constraint Manipulation and Visualization in a Multimodal Dialogue System
Proceedings of the ISCA Workshop on Multi-Modal Dialogue in Mobile Environments, Kloster Irsee, Germany, 2002.

Paper X Gustafson, J. and Sjölander, K.
Voice Transformations For Improving Children's Speech Recognition In A Publicly Available Dialogue System
Proceedings of ICSLP 02, vol. 1, pp. 297–300, Colorado, USA, 2002.

Acknowledgements

First of all I wish to thank my supervisor Björn Granström and my assistant supervisor Rolf Carlson for their support and guidance, and for giving me the opportunity to pursue all of my research interests. I am especially thankful to Rolf for introducing me to the dialogue field in 1992. He gave me a flying start by showing me how his dialogue manager in the Waxholm system worked, and by putting me in contact with international researchers in the dialogue field.

I am thankful to Jonas Beskow and Kåre Sjölander for providing me with excellent speech technology components, without which the dialogue systems would not have been possible to build. I thank Magnus Lundeberg for making August so handsome and for giving him life with funny and informative facial gestures. I would like to thank Johan Boye for a really fruitful and enjoyable collaboration on the AdApt and Pixie systems. I am grateful to Mats Wirén for many fruitful research discussions and for being able to always give wise advice.

I am filled with a warm happy feeling when I think about my collaborative work with Nikolaj Lindberg. Our wild and inspiring discussions and interesting co-operation made the development of August a great pleasure. I would like to thank Linda Bell for leaving the language-learning field and instead analyzing the August dialogue corpus. Since then we have had a very jolly and productive collaboration, regardless of where we have been working, and Linda has become a very dear friend. I would like to thank Jens Edlund for being a great friend and an invaluable colleague. Many great ideas have emerged (and vanished?) on late nights at Östra Station. I am very grateful to Eva Gustafsson for all her loving and support.

I am grateful to Anders Lindström and Leif Bengtsson at Telia Research for making it possible for me to finally write the thesis summary by giving me invaluable time and support. I would like to thank Linda and Nikolaj for their continued help and hard work during the whole process of writing the thesis summary. You made it more fun to finally pull myself together and to wrap it all up. I would like to thank Johan Boye, Mats Wirén, Anders Lindström, Robert Eklund, Arne Jönsson, Rolf Carlson, Eva Gustafsson, Mattias Heldner and Jean-Claude Martin for reading and commenting on drafts of this thesis. I would like to thank Marion Lindsay for helping me to increase the readability of the thesis for native speakers of English. All faults that remain are deliberate – to keep you awake and alert!

Thanks to all my former colleagues at TMH who made my eight years there very pleasant. I would like to thank my current colleagues at Telia Research for an inspiring environment, with everything from inhaling monkeys to WCDMA for UMTS. I would also like to thank Morgan Fredriksson and his colleagues at Liquid Media for a very enjoyable collaboration in the Pixie and NICE projects.

Finally, I am grateful to my family, especially my father Ragnar who has always supported me, and without whom I would never had made it this far.

Part of this work was carried out within the Centre for Speech Technology supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

PROLOGUE

"The Talking Robot was not a useful machine, and was only built for show. It only answered certain questions. Once an hour, a group of tourists stood before it. They told the robot engineer their questions. Then the engineer decided which questions to ask the Talking Robot. It wasn't very interesting. The robot could only answer easy questions like, 'How hot is the room at the moment?' or 'What is 25 x 459?' But you really don't need a robot for that. Few people came back to the Talking Robot a second time."

A description of a future attraction at a New York Museum in the year of 1998, from Isaac Asimov's story "Robbie" in the book *I, Robot (1950)*. The first version was published as "Strange Playfellow" in *Super Science Stories*, September 1940, pp. 67–77.

Chapter 1

INTRODUCTION

This thesis describes work carried out during the last ten years, aiming at developing multimodal spoken dialogue systems where users can express themselves freely without having to learn a special way of speaking. In all these systems the users have interacted in spoken Swedish with animated talking characters. To be able to develop these systems, human–computer dialogues have been collected and analyzed. The purpose of the studies has been to increase our understanding of how dialogue systems can utilize the regularities found in human–computer interaction.

1.1. Research issues

The general research aim has been to design multimodal spoken dialogue systems that allow users to communicate naturally and efficiently. For this purpose, two interrelated goals have been pursued:

1. To develop a series of multimodal spoken dialogue systems that would serve as experimental test benches.
2. To perform empirical studies of how users behave and interact with these experimental systems. "Users" have not only been subjects in a controlled laboratory setting but also people of different ages and backgrounds, who have interacted with these systems in public environments.

The user studies have provided guidance and inspiration for the next design iteration, and each successive dialogue system has in turn allowed for novel experiments and data collection.

To be able to study a wide range of user behavior, systems in a number of different domains have been implemented and used to collect human–computer dialogues. Five different systems will be presented in the thesis: The *Waxholm* system where users could ask about the boat traffic in the Stockholm archipelago; the *Gulan* system where people could retrieve information from the Yellow pages of Stockholm; the *August* system which was a publicly available system where people could get information about the author Strindberg, KTH and Stockholm; the *AdApt* system that allowed users to browse apartments for sale in Stockholm and the *Pixie* system where users could help an animated agent to fix things in a visionary apartment publicly available at the Telecom museum in Stockholm. All systems are the result of collaborative projects. Four of the systems were used to collect spoken dialogue corpora: Waxholm, August, AdApt and Pixie. The Gulan system was used for educational purposes. All systems will be described in chapter 4.

Apart from making it possible to pursue the two general aims presented above, the work of collecting and analyzing spoken human–computer interaction also led to the emergence of more specific research issues, e.g.:

- *How are subjects influenced by written scenarios?*
  In the Waxholm user experiments the subjects were found to reuse large parts of the written scenarios they were given. This was handled by adding a graphical representation of the domain as well as a multimedia introduction to the fully automated Waxholm system. This will be described in *Section 4.1.1.*

- *How are users influenced by the wording of the system output?*
  Paper II describes how subjects who interacted with a simulated system reused parts of the system questions in their answers. *Paper V* reports on an experiment with the simulated version of the AdApt system. In this study it was investigated if it would be possible to influence the users' choice of modality in their input by using a certain modality in the system output. Examples of verbal convergence in the fully automated AdApt system are given in *Section 4.1.4.*

- *How do users change their way of speaking when a dialogue system fails?*
  Analyses of the August dialogue corpus revealed some of the strategies people employ for error handling. When the users repeat a misunderstood utterance they modify their speech either by using other words in the repetition or by modifying the pronunciation towards a clearer articulation. A detailed analysis of this can be found in *Paper III* and *Paper IV.*

- *How does a dialogue system with an open microphone affect users' input?*
  The multimodal system AdApt used speech detection instead of a push-to-talk button. This led to fragmented utterances when the subjects took the turn by referring to an object on the screen, or by giving feedback on the system's previous turn. They would in many cases pause for a moment while considering what to say next. The initial feedback fragments are analyzed in *Paper VI.* To be able to handle these fragmented utterances, a new system architecture and a parser were developed. This allowed the system to wait for more input, if it regarded the user utterance as incomplete in the current dialogue context. An I/O handler that handled the timing of the multimodal input and output was added. The method for handling fragmented utterances is described in *Paper VIII.*

- *How should a system that allows for advanced turn-handling be able to communicate to the user whether it is waiting for more input or not?*
  The August, AdApt and Pixie systems used visual feedback for turn-taking. The animated face was used to encourage the users to keep talking. In the AdApt system, facial feedback was accompanied by icons intended to represent the relevant parts of the recognized utterances. The

turn-taking gestures used in the AdApt system are shown in *Section 4.1.4.* and the icon handler is described in *Paper IX.*

- *What happens when you put a spoken dialogue system with multiple domains in a public environment?*
  Experiences from the August and Pixie systems showed that people are inclined to engage in a socializing dialogue where they talk about the context of the dialogue, e.g. about the agent, the exhibition or the previous discourse. Furthermore, it is possible to influence the users to talk about topics that the system can handle. This will be described in *Section 4.4.* and is also discussed in *Paper IV.*

- *How can recognition of children's speech be improved, when only acoustic models trained on adult speech are available?*
  Many children interacted with the August and Pixie systems. *Paper III* deals with how children and adults modify their pronunciation during error handling. The effects of these modifications on the KTH speech recognizer are also discussed. The Pixie system included a commercial speech recognizer trained on adult speech, with telephone bandwidth. To decrease error rates, the children's voices were acoustically transformed on the fly, before being sent to the recognizer. Learning from the difficulties of assessing gender and age of speakers in the August corpus, the users of the Pixie system had to provide this information before interacting with the system. *Paper X* presents details on the voice transformation method and the result of using it.

1.2. Thesis overview

The thesis consists of two parts. The first part is an introduction to the field of spoken dialogue systems, followed by a short description of the thesis work. The second part consists of ten internationally published scientific papers. The thesis is outlined as follows: In Chapter 2, speech interfaces are introduced and compared to graphical interfaces. The advantage of combining them into multimodal interfaces is also discussed. Finally, the point of embodying the speech interface is reviewed. Chapter 3 introduces spoken dialogue systems. It describes how spoken dialogue systems are developed and what they can be used for. Chapter 4 describes the five dialogue systems discussed in this thesis and gives some examples of specific research issues that they have highlighted. It also describes the different system architectures of the implemented dialogue systems. Finally, it describes some relevant features of the systems and the settings in which they were used to collect human–computer dialogues. An overview of the topics of the included papers is presented in Chapter 5, and Chapter 6 summarizes some of the findings in the thesis work. The second part of the thesis contains the ten research papers that make up the basis for this thesis.

Chapter 2

BACKGROUND

Today, users mostly interact with computers via direct manipulation in Graphical User Interfaces (GUIs). The work presented in this thesis aims at providing computer systems with speech interfaces as well. This chapter summarizes some of the advantages of using speech in human–computer interfaces. It also argues for combining speech and graphical interfaces into multimodal interfaces. Finally, it discusses the value of embodying speech interfaces.

2.1. Speech interfaces and graphical interfaces

The literature provides a number of reasons for using speech in human–machine interaction, some of which have been summarized by Cohen (1992) and Cohen and Oviatt (1995). An obvious advantage is that you can speak without using your hands and that you do not have to turn your attention to a computer screen. This feature makes speech as an interface especially suitable for people who do not see well or cannot move their hands easily (Damper 1984). The hands/eyes free property is also useful in situations where these resources are used for other tasks, e.g. data entry and machine control in factories (Martin 1976). Using speech instead of a keyboard in these situations can reduce error rates. Nye (1982) reported that a speech interface for supplying the destination of baggage at an airport produced less than 1% errors, compared to 10% to 40% for keyboard input. Another hands/eyes busy situation is driving a car (Julia and Cheyer 1998, Westphal and Weibel 1999). Nowadays car drivers can choose to operate mobile phones, navigation systems and advanced information systems. Speech control of these is safer than using a graphical interface, since the driver does not have to turn his attention from the road to the interface to control it by hand. However, it is important to design these new speech controlled systems with care so they do not overload the driver with more tasks than can be handled. The systems could for example be dependent on the driving situation, so that they keep quiet in situations where the users need to focus on the traffic.

In small devices with limited screen size and keyboards, graphical interfaces can only be used for simple tasks with a small set of actions. Hence, speech interfaces are especially advantageous for small mobile devices. They are also useful for large-scale displays or virtual environments (Julia et al. 1998, Pavlovi et al. 1998). A coming trend is to embed the information technology in the environment, removing the screen altogether. The EC Information Society Technologies Advisory Group has presented a future concept called *Ambient Intelligence*, which combines Ubiquitous Computing with Intelligent User Interfaces (Ducatel et al. 2001). According to their vision, information technology is present everywhere, but without being visible or imposing. Interfaces should appear when needed, and then be easy to use, context dependent and personalized. Speech and gesture recognition are among the key technologies identified as necessary to be able to realize this vision.

Speech communication is an efficient way of transmitting information between humans, who have communicated using spoken language for thousands of years. However, Noyes (2000) questions whether the situation of talking to a computer can be regarded as natural. An often-claimed advantage of spoken human–computer interfaces is that since they are natural for humans they would be universally accessible. According to Buxton (1990) natural does not mean universally accessible, at least not without having to be learned first. Natural languages like conversational English and German differ both in vocabulary and syntax and they can be regarded as natural for speakers that have acquired fluency in using them. Thus, having to learn a syntax and vocabulary that is appropriate for the tasks that are related to a specific domain does not make a speech interface unnatural, even though it makes it less universally accessible. If a speech interface is to be regarded as natural, it must be obvious how to express the desired concepts of the domain and the users have to be able to express themselves in a rich and fluent manner. Usually when humans communicate with computers they interact via GUIs where they receive information visually, and input information via a keyboard or pointing device. This kind of interaction is not necessarily natural either as was exemplified in the following scene from a Star Trek movie, that is shown in Figure 1 on the next page.

> Scotty, having been transported from 2200 back through time to the late 1980s, attempts to use a Macintosh computer. At first, he speaks to the Mac from across the room:
>
> Scotty: *Computer! – Computer?*
>
> His friend Bones quickly realizes that the primitive 1980s technology does not respond directly to voice commands, so he hands Scotty a mouse. Scottie takes the mouse and then holds it up to his mouth like a microphone saying:
>
> Scotty: *Ah! - Hello computer!*
> Technician: *Just use the keyboard!*
> Scotty: *The keyboard? How quaint!*

Figure 1. *A transcript from the film Star Trek IV: The Voyage Home (1986).*

In direct manipulation interfaces, users interact by selecting linked texts or icons that represent commands to the system. A limiting factor is that everything the users want to do at any given time must be represented in the GUI. To overcome this limitation many GUIs also make all commands available through keyboard shortcuts. However, the meanings of the words used in menus, the icons in the tool bars and the keyboard shortcuts all have to be learned by the users. This would not be necessary in a system where the users could say what they wanted to do using unrestricted spoken language.

Spoken interaction can be faster if users immediately can say what they want to achieve without going through the menus or hierarchical pages that are used in GUIs. Users can give a number of information units in one single utterance, e.g. saying *I want to go from Stockholm to Waxholm today at about five o'clock* instead of selecting a number of popup menus in a GUI. If you want to build more intelligent systems, natural language makes it possible to construct complex messages that would be hard to input graphically, e.g. *Why is this apartment more expensive than the one downtown that you showed me before?* Furthermore, users can communicate their attitudes and emotions simultaneously by providing their verbal message with certain prosodic cues. These can be used in dialogue systems to detect if something has gone wrong in the previous discourse (Hirschberg et al. 2000) or to detect self-repair in spontaneous speech (Nakatani and Hirschberg 1994).

However, the freedom and efficiency that speech gives users also makes speech harder for the computer to handle. In spoken interfaces the users can at any time choose to say whatever they want regardless of what the dialogue designer had anticipated. In a spoken dialogue system a user who is posed with a question might answer with a meta question, *Please state your security number – Why do you want me to do that?*, a rejection *Please state your security number – Forget it!*, with a clarification question, *When do* *you want to leave – What tickets are available?* or with a related request, *When do you want to leave – I want to take the express train!* Thus, it is important to conduct studies with real users to be able to anticipate how they will react when performing tasks using spoken dialogue systems. In GUIs this is not a problem since the users are limited to the actions the interface designer has decided should be possible to perform. The system will for example not continue until the users press the OK button in certain contexts. Moreover, objects and actions are visually represented in the interface and there are a limited number of ways for users to express what they want to do, e.g. deleting a file either by selecting a file icon and pressing delete or by dragging the file icon to a trashcan icon. This makes it possible for users to explore the possibilities of the system and it makes it easy for the system to understand what they wish to do. However, the simple syntax that is used in GUIs limits the types of tasks they can be used for. Speech interfaces of today can handle more complex syntax, but they cannot understand unrestricted spoken language – to get a reasonable performance they have to use a restricted dictionary and grammar, which leads to a vocabulary problem. It is hard for users to know the limitations of what they can say, and to explore the set of possible tasks they can perform using speech (Yankelovich 1996). It is difficult for the dialogue designer to anticipate how people will express what they want to do. Furnas et al. (1987) found that even people interacting with computers via command language will use many different terms to express the same thing, and Brennan (1990) refers to a report from the HP Natural Language project 1986, called “7000 variations on a single sentence.” Nonetheless, there are some general features of spoken interaction that make it possible to predict what people will say when they engage in spoken dialogue. There are regularities in dialogues that can be used when designing spoken dialogue systems. People usually adjust their way of talking according to the receiver, hence also when they interact with computers. Computer-directed speech has in previous studies been shown to be simpler in syntax resulting in shorter utterances, has smaller lexical variation and uses ambiguous pronouns and anaphoric expressions in a restricted way (Guindon 1988, Kennedy et al. 1988, Dahlbäck 1991, Oviatt 1995, Bell and Gustafson 1999b). People also tend to use the same words as the system when referring to various concepts in the dialogue (Brennan 1996, Gustafson et al. 1997).

Another problem with speech is that it uses a lot of short-term memory (Karl et al. 1993) and takes up the linguistic channel, which according to Schneiderman (2000) makes speech interfaces less suitable for some types of complex tasks that also need the linguistic channel. Such a task could for example be to write a business text (Leijten and Van Waes 2001). Speech often requires planning during execution, which might lead to fragmented utterances (Bell et al. 2001) or disfluent speech (Oviatt 1995, Yankelovich et al. 1995). Users who interact with systems with high error rates are also often disfluent (Oviatt et al. 1998). Yet another problem to overcome is the fact that speech recognition is still quite error prone, which makes it less reliable than traditional graphical interfaces. According to Schneiderman (1997) a user interface must do what the users intended it to do otherwise they will lose confidence in the system and stop using it. It must also be possible for the users to inspect which command the system received from the users before it is executed, giving the users a feeling of control. This is possible in a GUI, but hard in speech interfaces since speech is dynamic and volatile. It is possible to use verbal confirmations of what the system understood in each turn, but as Boyce (1999) points out people will regard such a system as slow and tedious.

Speed is a problem for speech on the output side of a system, since the information has to be conveyed serially piece by piece. In GUIs a lot of information can be presented at the same time, which makes it possible for users to browse or skim through the information to get an overall feel of the material and then access the interesting parts more carefully. Large amounts of structured information is for example often easier to convey graphically in a table than verbally in a spoken dialogue system – especially if the users have to compare a number of features between a limited number of objects. However, if there are many features and a very large number of objects an intelligent spoken interface could be better. Carefully designed, it could guide the users to the most relevant objects and help the users to interpret the difference in features between objects.

This section has presented a number of advantages of speech interfaces, but also a number of challenges in dealing with spoken user input. Instead of arguing about which type of interface is the superior one, it would be more interesting to investigate how they can be combined into a multimodal interface. The next chapter deals with how spoken and graphical interfaces can be integrated, and it gives examples of advantages and problems of multimodal interfaces that have been found in different studies.

2.2. Multimodal interfaces

Spoken and graphical interfaces have their respective benefits, which means that it could be advantageous to let an application use both, and let the users change modality depending on the situation. For example, an email application that is used on a desktop computer at the office should probably be provided with a graphical interface. However, when it is used on a small mobile device in the car it would probably be better with a spoken interface. But then again, if the user drives past a construction site with a lot of noise, or if the user goes by public transportation and wants privacy, a GUI would be preferable. The solution would be to have both a spoken and a graphical interface and let the users decide for themselves which modality they prefer on different occasions.

Systems that use more than one channel/modality to communicate information are called either multimedia or multimodal systems. The difference is that multimodal systems use a higher level of abstraction from which they generate output and to which they transform the user input (Coutaz et al. 1994). This means that multimodal systems can render the same information through different output channels, and that they can fuse user input that was transmitted through multiple channels into a single message. A benefit of multimodality is the fact that users can combine multiple modalities to transfer a single message, which has been shown to decrease error rates (Bangalore and Johnston 2000, Oviatt and VanGent 1996). According to the TYCOON framework six basic types of cooperation between modalities can be defined (Martin et al. 1998, Martin 1998):

Equivalence Several modalities are suitable for transmitting the same information.

Specialization Some modalities are better than others for transmitting the information, e.g., it is better to convey spatial information in a visual than a verbal channel.

Redundancy Exactly the same information is transferred through multiple channels at the same time, and this could e.g. be used to prevent speech recognition errors.

Complementarity Several modalities are used together to convey the information, e.g. selecting an icon using the mouse while saying: *How much does this one cost?*

Transfer Information that was produced by one modality is used by another modality. An example would be to use mouse input to restrict the grammar of the speech recognizer.

Concurrency Several modalities transfer independent information units at the same time, e.g. using a voice command to save a document that is being keyboard edited in a word processor.

If two modalities are equivalent users can switch between them to avoid and correct errors. Oviatt (1992) showed that people who could use either speech input or pen input in a system, switched to pen input when entering foreign names and alternated between modalities to resolve repeated errors.

According to Grasso (1997) speech and direct manipulation have different specializations that it would be beneficial to take advantage of when building human–computer interfaces. GUIs are good at handling a few and visible references while speech interfaces are good at handling numerous and non-visible references. GUIs handle simple actions very well but cannot handle the complex actions that speech interfaces make possible. Furthermore, graphic representation is persistent in contrast to speech which is non-persistent. This feature was used in the AdApt system, presented in Paper IX, where graphical icons were used instead of verbal confirmation to give feedback on what the system thought the users had asked for. In addition to this, the current set of search constraints specified so far in the dialogue was visualized, making it possible for the users to inspect and change previously given constraints.

Multimodal redundancy does not seem to be very common. Petrelli et al. (1997) report that people who used their multimodal system rarely transmitted redundant information through multiple channels. Redundancy can be used to ensure that the information is correctly understood, e.g. in noisy environments or during error resolution. However, Oviatt (1999) only observed 1% redundant multimodal commands during error resolution. Oviatt et al. (1997) reported that people use modalities in a contrastive manner to communicate a shift in content or functionality. Similarly, in the AdApt user studies only a few examples of redundant multimodal input were observed. In this system apartments were indicated as colored squares on a map and these could be selected with mouse input. Some users would select apartments graphically when they shifted focus from one apartment to another, even in cases when they referred to it verbally. This resulted in partly redundant multimodal input like when a user clicks on the red square while saying *How much does the red one cost?*

Complementary use of several modalities is the most common multimodal pattern. Oviatt et al. (1997) have shown that it is possible to use redundancy and complementarity between n-best lists for graphical and spoken input in order to get the correct interpretation of the multimodal command, even though none of the n-best lists had the correct interpretation as number one. It can be hard to decide if a combination of modalities is redundant or complementary. Martin et al. (2001) propose an axis of “salience values” where the combination is regarded as complementary if this value is zero and very redundant if it is one.

Transfer can be used to improve speech recognition by limiting its grammar according to mouse clicks. Bangalore and Johnston (2000) used finite-state transducers to allow the gestural part of multimodal utterances to directly influence the speech recognition search, thus reducing the error rates by about 23%.

According to Martin's definition, modalities are concurrent if they are independent of each other, but used in the same system. Even in cases where they are used simultaneously, their input should not be merged. However, simultaneous input from several modalities has rarely been found at all, not even in cases where they should be merged. This applies for example for systems with spoken and graphical interfaces where the graphical input normally precedes the verbal input. Oviatt et al. (1997) reported that about 25% of all multimodal commands were concurrent. Typically, users would submit the graphical part of the command between one and two seconds before the verbal part. In the AdApt system concurrent multimodal input was only found in some rare cases (Gustafson et al. 2000). This is of course dependent on the applications and the kind of input devices that are used for gesture input. Future multimodal systems might elicit more concurrent commands. If the system could interpret the users' hand gestures and facial expression in a visual modality, it might for example be natural for these to occur concurrently with the speech. However, this remains to be verified in experimental studies.

There are a number of possible output modalities that systems can use, e.g. recorded or synthesized speech, non-speech sounds, written text, graphs, maps, tables or embodied characters that use gestures and facial expressions. Input modalities could for example be speech, pointing and gestures in 2D or 3D, characters or hand-writing, eye movements, lip movements, facial expressions or keyboard and mouse input (Benoît et al. 2000). Bernsen (2001) presents taxonomies of input/output modalities as well as a methodology that can be used to select the most useful combination of input/output modalities for a certain application.

One modality that humans use while speaking to one another is the visual modality of facial and body movements. The next section describes how embodied conversational agents can be added to dialogue systems, resulting in systems with multimodal spoken output.

2.3. Embodied interfaces

Humans who engage in face-to-face dialogues use non-verbal communication such as body gestures, gaze, facial expressions and lip movements to transmit information, attitudes and emotions. If computers are to engage in spoken dialogue with humans it would seem natural to give them the possibility to use non-verbal communication too. An embodied conversational character could increase the believability of the system and make the interaction more natural. Previous studies have shown that users who interact with an animated talking agent spend more time with the system, enjoy the interaction more and think that the system performed better. This has been called the *persona effect*, and it is considered by many researchers to be the most important reason for adding animated agents in educational systems (Walker et al. 1994, Koda and Maes 1996, Lester et al. 97, van Mulken 1998, Lester et al. 1999). There is a risk that the interaction becomes slower when users try to interpret all the signals that the face emits, even though they were not deliberately inserted by the interaction designer (Takeuchi and Naito 1995). This means that some types of animated agents might distract the users from their tasks (Koda and Maes 1996, McBreen and Jack 2001). However, Pandzic et al. (1999) and Walker et al. (1994) did not find any degraded task performance when using embodied agents.

Another concern is that embodied agents will lead people to anthropomorphize the interface, resulting in too high expectations of the intelligence of the system (Takeuchi and Naito 1995, Koda and Maes 1996, Walker et al. 1994). On the other hand, Reeves and Nass (1996) have shown that users tend to interact socially with computers in the same way as they interact with people even though the system does not have a human appearance. Laurel (1990) and Cassell et al. (1999) argue that interface designers could take advantage of anthropomorphism by embodying some types of interfaces, thus making the interaction more natural. Dehn and van Mulken (2000) have reviewed a number of studies on the usefulness of animated characters, and they conclude that most of these studies have failed to show an increase in user performance. Nevertheless, they argue that most of these studies were conducted on too short sessions, and that it would be desirable to do user studies on longer and multiple sessions. The animated agents' entertaining features could for example be used to motivate students to interact with educational systems. They believe that if animated characters are used correctly, larger studies will yield better user performances.

2.3.1. Facial appearance

Adding a face can make the dialogue situation more entertaining and engaging. The appearance of the face communicates who the speaker is by means of personality, social status mood, etc. This could be used in dialogue systems to increase the users' trust and satisfaction (Nass et al. 2000). Most humans are very good at recognizing and remembering faces (Donath 2001), a feature which can be used to make different speech services memorable and familiar. The appearance of the agent can be used to communicate the system domain. This can be done using a famous real or fictive person's face or by dressing the characters to show that they belong to a certain occupational group. If a single dialogue system supplies a number of different services, domain specific recognizer lexicons and dialogue managers could be loaded depending on which character the user is speaking to, e.g. load the food domain when they are talking to the Swedish chef and the sports domain when they interact with the virtual sports commentator. It could also be useful to have multiple characters with different personality within the same domain. André et al. (1999) describe a market place with a number of embodied characters that were given different personalities.

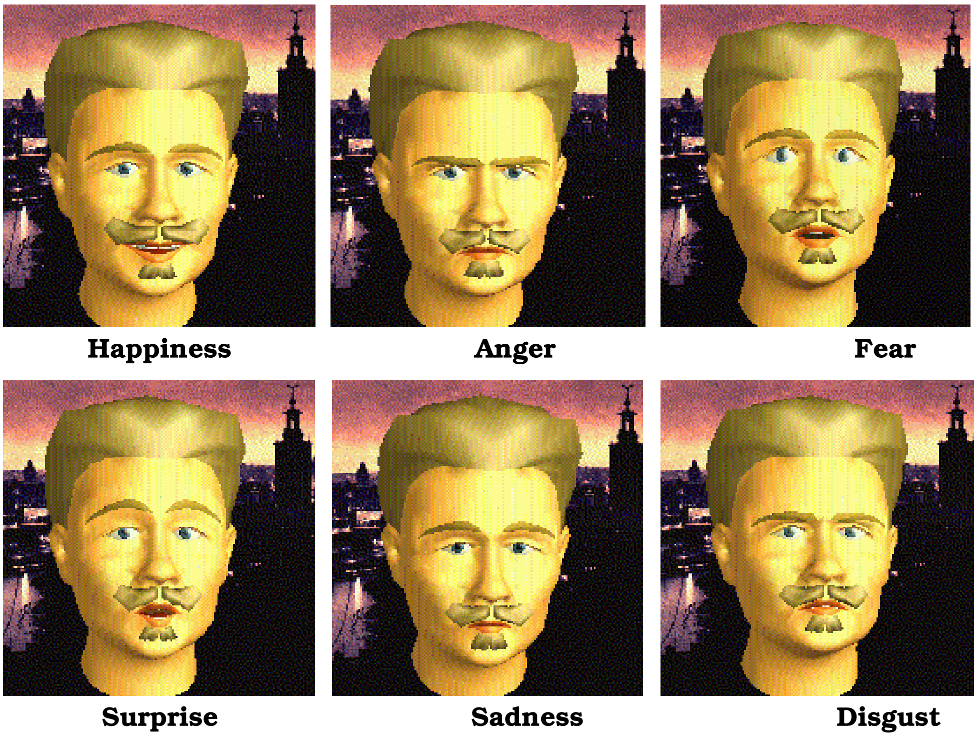
2.3.2. Facial gestures

Animating the face brings the embodied character to life, making it more believable as a dialogue partner. According to Ekman (1979) facial actions can be clustered according to their communicative functions in three different channels: the phonemic, the intonational and the emotional.

The phonemic channel is used to communicate redundant and complementary information in what is being said. Fisher (1968) coined the term *viseme* for the visual realization of phonemes. Accurate lip movements in audiovisual speech can improve intelligibility, especially for the hearing impaired (Agelfors et al. 1998), but also in general in noisy environments (Benoît et al. 1994, Beskow et al. 1997). To be able to produce 3D animations of audiovisual speech, appropriate face models have to be developed. These models can be either physically based like Waters' model (Waters 1987) or parametric like Parke's model (Parke 1975). The Parke model has been used in several audiovisual speech synthesis systems (Lewis and Parke 1987, Cohen and Massaro 1993, Beskow 1995). There are also 2D facial animation systems that use image processing techniques to morph between recorded visemes (Bregler et al. 1997, Ezzat & Poggio 1998).

The intonational channel is used to facilitate a smooth interaction. Facial expressions, eyebrow raising and head nods can be used to communicate the information structure of an utterance, for instance stressing new or important objects (Scherer 1980, Pelachaud et al. 1994, Cassel et al. 2001, Decarlo et al. 2002).

The emotional channel is used to increase the animated character's social believability. Ekman et al. (1972) found the six universal emotions that are interpreted by August in Figure 2 (Lundeberg and Beskow 1999). There are *display rules* that regulate when speakers show emotions. These rules depend on the meaning the speaker wants to convey, the mood of the speaker, the relationship between speaker and listener and the dialogue situation (Ekman 1982). Some animation systems have implemented such display rules (Poggi and Pelachaud 1998, de Carolis et al. 2001). Cassell and Thórisson (1999) found that adding gestures for dialogue regulation, i.e. turn-taking gestures, in their Ymir dialogue system increased user satisfaction more than it did when adding emotional gestures. Guye–Vuillieme et al. (1999) argue that the domain of Ymir (the solar system) had little emotional content and they conclude that both kinds of feedback are needed to get more user-friendly virtual environments.



Happiness **Anger** **Fear**

Surprise **Sadness** **Disgust**

Figure 2. *Ekman's universal emotions, as interpreted by August.*

2.3.3. Body gestures

Rimé and Schiaratura (1991) have presented a classification system with six classes of gesture usage in dialogues. *Speech markers (beats, batons)* are used to communicate the information structure of an utterances, e.g. to stress important or new objects in a verbal utterance. *Ideographs* are produced while the speaker is preparing an utterance to indicate the direction of thought. *Iconic gestures* are used to show some representation of an object that is being referred to verbally. The gesture can depict the shape, some spatial relation or action of an object. *Pantomimic gestures* play the role of the referent. *Deictic gestures* are used to point to objects visual in the users environment or represented in the graphical interface. Finally, *Emblematic gestures* are gestures that have a direct translation into words that is known in a specific culture or social group. They are used to send messages like thumbs up for "ok", which is shown in Figure 3 among other examples of gestures used in the Pixie system.

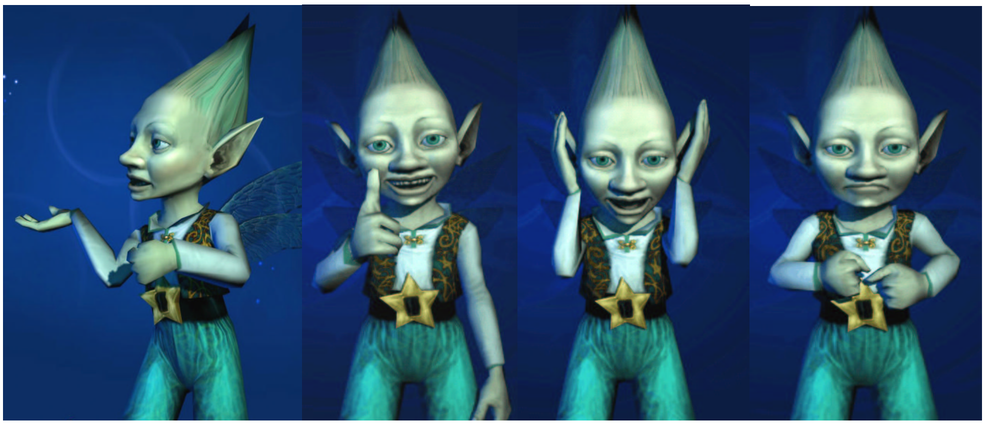


Figure 3. *Some of Pixie's body gestures (Liquid Media 2002).*

2.3.4. Gaze

According to Kahneman (1973) gaze indicates three types of mental processes: spontaneous looking, task-relevant looking and looking as a function of orientation of thought. Thus, in conversation gaze carries information about what the interlocutors are focusing on. Gaze can be used to communicate the speaker's degree of attention and interest during a conversation, to regulate the turn-taking, to refer to visible objects, to show the speaker's mental activity, to display emotions or to define power and status. Pelachaud et al. (1996) described a facial animation system that among other things could display different gaze patterns. According to Duncan (1972) speakers can give cues that indicate the end of their turns not only with prosody and syntax, but also by changing the direction of their gaze. According to Goodwin (1981) the listener looks away from the speaker while taking the turn to avoid cognitive overload while planning what to say. The usefulness of gaze in turn-handling was investigated by Cassell et al. (1999). They found that the speakers looked away from the listeners at the beginning of turns and towards the listeners at the end of turns. They also found that speakers tended to look away from the listeners while giving old information (theme) and towards the listeners while giving new information (rheme). If theme coincided with the start of a turn, the speakers always looked away from the listeners. Thórisson (2002) describes a turn-taking model called the Ymir Turn-Taking Model (YTTM) that uses speech detection, prosody, gesture and body language to determine when the animated agent should take the turn. The BEAT system uses gaze, head nods and eyebrow-raising for turn-handling (Cassel et al. 2000). Finally, according to Colburn et al. (2000) turn-handling gaze can be used to indicate who is talking in multi-party dialogues such as virtual conferencing.

Chapter 3

SPOKEN DIALOGUE SYSTEMS

In spoken dialogue systems the users' spoken input is translated into computer readable form by a speech recognizer (ASR). The output from the recognizer could be orthographic words, syntactic classes or application specific commands that occur in sentences, n-best lists (lists of possible sentences) or hypothesis lattices. The output from the recognizer is sent to a linguistic understanding component that interprets the semantic meaning of the input, which in turn is used by the dialogue manager to determine what to do, e.g. perform a database search, send a command to an external device or ask a clarification question to the user. The system also communicates with speech output, using either recorded prompts or speech synthesis.

To date, the speech recognizer and the linguistic understanding components have had to use limited lexicons and grammars in order to get reasonable performance. However, in some services with simple dialogue structure and where it is possible to collect large speech corpora, statistical grammars can be built that have less limited coverage. An example of such a service is call routing, where the system sends an incoming telephone call to the appropriate operator (Arai et al. 1998).

At every given point in a dialogue either the system or the user has the initiative. If the same part controls the dialogue all the time it is called single initiative, while it is called mixed initiative when the initiative changes over time. If the task model determines who has the initiative it is called fixed mixed, and if both dialogue partners can take the initiative at any given time it is called dynamic mixed (Allen 1997).

Most commercial spoken dialogue systems use system initiative, where predefined slots are filled or where the users are prompted with menu choices. In these systems the structure of the application determines the structure of the dialogue. While menu dialogue systems are appropriate for many simple tasks, they are not suitable for large vocabulary applications or for applications where the users have to provide the system with a lot of data (Balentine 1999). It is problematic to build large and complex applications since menus preferably should not contain more than about five items (Balentine & Morgan 1999, Garder–Bonneau 1999) and because deep menu structures should be avoided (Virzi & Huitema 1997). Moreover, since the menu hierarchy is built from the structure of the backend system, users are required to know how the system is organized in order to be able to find adequate help.

In contrast, in a system with dynamic mixed initiative users can say what they want to do without having to learn a special way of speaking, and without knowing the organization of the backend system. However, since such dialogues are not strictly system driven it is more difficult to understand the underlying intention of the users' utterances. User adaptive spoken dialogue systems cannot be built without studying both human–human dialogues and human–computer dialogues. To be able to study human–computer dialogues both real and simulated systems have to be developed. By studying human–human interaction it is possible to take advantage of the rules and regularities that it reveals. Furthermore, it is very important that conversational systems are able to handle errors and try to prevent them from occurring, by communicating what has been understood and if necessary initiate a clarification dialogue to solve communicative problems. This also requires the collection and study of user data.

3.1. System architectures

Spoken dialogue systems usually have three parts: Understanding the user input, deciding what to do, and generating the system output. In simple question/answer dialogue systems this can be done in a pipeline manner where one module sends its output to the next module and finally an answer is generated, as seen in Figure 4.

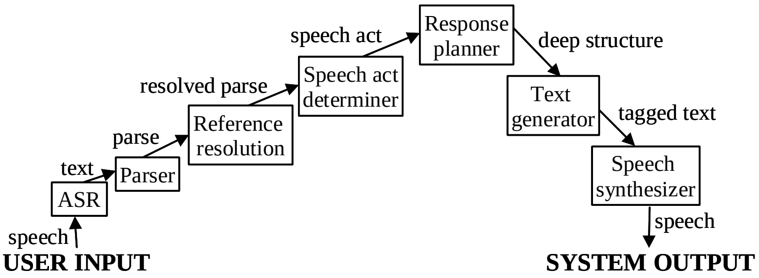


Figure 4. *A system architecture for a simple spoken dialogue system.*

If the system is to be multimodal and conversational a more complicated system architecture is needed. The system must be able to combine input from several modalities, which means that it in some cases has to wait for more information from the same or another channel before sending the input to the dialogue manager. To make the system reactive it has to be able to produce output while it is processing the input, for example producing turn-handling facial gestures while listening to speech input. Finally it has to be able to decide which channels to use for output and when to produce it. Figure 5 below shows an example of what an architecture that can handle some of these issues might look like. Apart from the vision input, this architecture is almost identical to the architecture used in the AdApt system. Another difference is that the I/O Manager has been divided into three sub-modules, one module for merging input from different input modalities, one module for decomposing multimodal messages from the dialogue manager that are to be sent to respective output module, and a module that is responsible for the timing of the input and output of the system.

In order to build conversional systems it is important to be able to handle user utterances that contain problematic parts, due to either recognition errors or user hesitation or disfluencies. The system also has to respond fast to give the dialogue a conversational feel. The demands that conversational systems put on the understanding modules are quite hard to meet. The system should be able to understand the intentions of the user, use planning to decide what to reply and then answer very fast. Some of these requirements can be met by using machine learning for the semantic analysis.

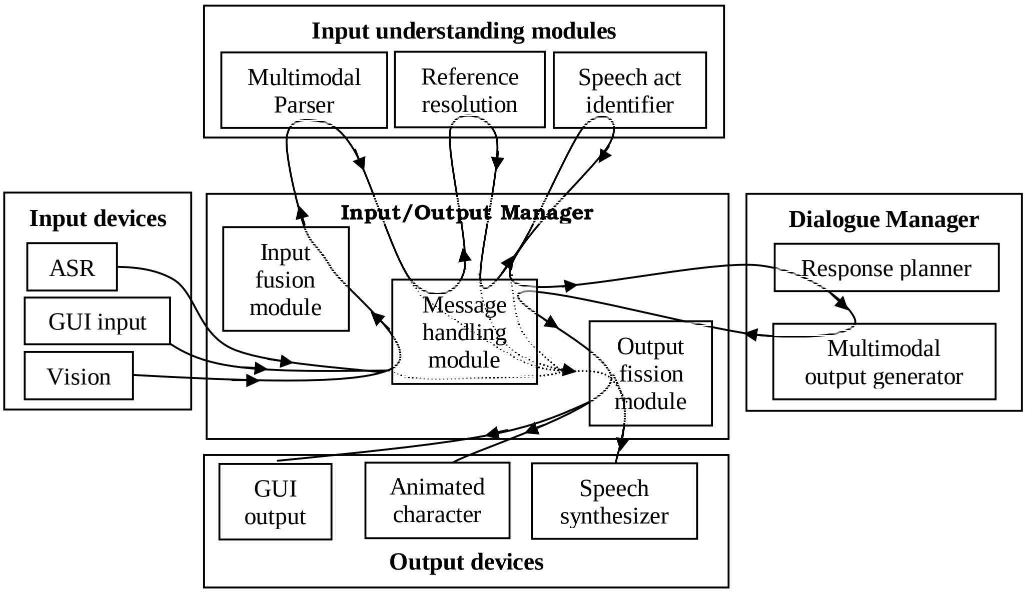


Figure 5. *A system architecture for a multimodal conversational system.*

Spoken dialogue systems are quite complex with many different components, which also puts constraints on the system architecture. Since many developers have to work together to build the systems it is necessary to make them modularized. This modularization could be done in different ways: either the system is built in an object-oriented language where the whole system could be run in one process and where there are different internal modules/objects that communicate via internal interfaces; or the modules could be distributed into multiple processes that communicate via external interfaces, e.g. sockets. The latter makes it easier to build a system that is distributed over several computers and more importantly it is possible to implement the different modules in different programming languages. This makes it easier to distribute the work of implementing the modules to developers with different backgrounds and requirements on the programming language. A drawback of the distributed architecture is that it might be slow in cases where a lot of information must be communicated at a high rate. It may also make the installation and maintenance more complicated.

The next chapter will present an overview of how to develop spoken dialogue systems. It will also give a short introduction to knowledge sources that have been used when developing such systems.

3.2. Building spoken dialogue systems

One difficulty when building spoken dialogue systems is that it is hard to anticipate how people will speak to the system. Furthermore, since the users' way of speaking will be influenced by the functionality of the system, it would be desirable to do user studies under realistic conditions before deciding on the design of the dialogue system. Wooffitt et al. (1997) present three solutions to the problem of predicting how users will interact with spoken dialogue systems. The first is *design by inspiration*, using the fact that humans are experts in human language. In these cases the application is analyzed and a strictly system driven dialogue system that is developed uses the linguistic intuition of the system designer. As Wooffitt et al. (1997) point out this is not a very good idea since the designer usually cannot think of all possible situations in advance. Another problem is that this method relies on the designer's linguistic competence and not his knowledge of language use. The next method is *design by observation*. This means that the designer observes how people solve the same tasks while talking to other humans. To be able to do this it has to be possible to collect human–human dialogues. If there is no manual version of the service this is of course impossible. In those cases it is necessary to *design by simulation*. This is the well-known Wizard-Of-Oz (WOZ) technique, where some or all parts of the system are simulated by a human operator. To get realistic user interaction it is important that the users believe that they are interacting with a real system.

Bernsen et al. (1998) presented a life cycle for the development of spoken dialogue systems, see Figure 6.

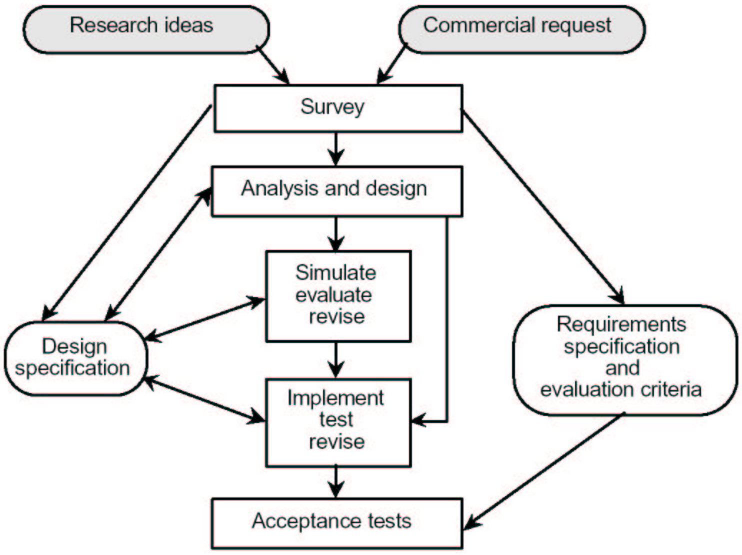


Figure 6. *The life-cycle from Bernsen et al. (1998).*

The life cycle starts either with research ideas or commercial requests that are used in a survey that aims at producing design specifications, requirement specifications and evaluation criteria. The design specification is first used to develop a simulated version of the system. This is exposed to test users and the evaluation of the user interactions is used to revise the design specification. Then a fully functional system is built, user tested and the design specification is revised iteratively until the requirement specifications are met. The evaluation criteria are then used to do acceptance tests with the end users.

There are a number of knowledge sources that are useful in the development of spoken dialogue systems. Here is a brief overview of three types of knowledge sources.

3.2.1. Human–human communication theories

Human–human dialogues have been studied extensively and there are many theories that aim at modeling different aspects of communication. This section will give some examples of theories about human–human dialogue that have been influential for designers of human–computer dialogue systems. Before going into these theories the usages of the term conversation will be commented on.

Conversation and conversational

Humans use language to perform many communicative functions. Traditionally, spoken language mostly has had an interactional function - to establish and maintain personal relationships, while written language mostly has had a transactional function - to transfer information (Brown and Yule 1983). This is not completely the case anymore - people leave short messages verbally on answering machines and they write e-mails and sms-messages to maintain their personal relationships. According to Leech et al. (1995) "*conversation ... is dialogue conducted primarily for interactional, rather than transactional reasons*", but others, for example Sacks et al. (1974) use the term conversation for any unscripted dialogic talk. Levinson (1983) points out the following about conversation "*conversation is not a structural product in the way that a sentence is - it is rather the outcome of the interaction of two or more independent, goal-directed individuals, often with divergent interests*". Button (1990) argues that even though it is possible to build machines that simulate conversational sequences, it would be wrong to say that they are "conversing" in the same way as humans. He claims that this has implications on how conversational analysis should be used when developing dialogue systems. Zue and Glass (2000) and Allen et al. (2001) use the term conversational dialogue systems to indicate that they allow the users to state what they want to do freely - just as they would if solving the task by talking with another human. However, the goals of these conversational human–computer interactions are still primarily task-oriented. There are a number of spoken dialogue systems that can be called conversational according to this interpretation of the expression, e.g. the *How may I help you?* system (Gorin et al. 1997), the *MIRACLE* system (Stein et al. 1997), the *Jupiter* system (Zue et al. 2000), the *August* system (Gustafson et al. 1999) and the *AdApt* system (Gustafson et al. 2000).

Speech acts

Speech act theory is based primarily on the works of Austin (1962) and Searle (1969). The speech act theory deals with the communicative function of utterances, i.e. the intention of the speaker and the effect on the listener. It is highly relevant when designing spoken dialogue systems, since for each user utterance the system must decide its purpose: whether it is a request for information, a clarification question, a confirmation, an action, a change of topic, etc. It can be hard to assign speech acts to utterances in dialogue systems, since the same utterance can be associated with multiple speech acts depending on a range of factors, such as prosody and dialogue context. A number of plan-based computational dialogue models which use speech acts as plan operators have been developed (Cohen and Perrault 1979, Allen and Perrault 1980, Cohen and Levesque 1990, Litman and Allen 1990, Carberry 1990, Lambert 1993, McRoy and Hirst 1995).

Conversational structure

The main feature of a dialogue that distinguishes it from a monologue is that there are at least two partners who contribute to the discourse. This feature has been called the “chaining principle” (Good 1979). The dialogue consists of turns that are composed by smaller so called turn construction units (TCUs). These are potentially complete turns, which means that at the end of a TCU it is possible but not obligatory for the listener to take the turn. These places are called transition relevance places (TRPs). Turns can have various components, from a single phone to several utterances (Sacks et al. 1974, Schenkein 1978). The overlap in speech between interlocutors is less than 5%, while at the same time the silent intervals between turns are typically only a few tenths of a second (Levinson 1983, Ervin-Tripp 1979). Bull (1996) found that a third of the between-speaker intervals were less than 200 ms long - which is typically the shortest possible response time to speech. This means that the listener uses a range of features in the speaker’s speech to anticipate where the TRP will come. Studies on how speakers indicate and listeners perceive TRPs have for example found the following features to be relevant: cue words (Grosz and Sidner 1986), intonation (Hirschberg and Pierrehumbert 1986), boundary tones and silences (Traum and Heeman 1997), control phrases, topic and global organisation (Whittaker and Stenton 1988). In dialogues there are regularities in the ordering at a local level described as *adjacency pairs* (Schegloff 1968), for example Question–Answer. This simple structure is not always applicable, there is often an insertion-sequence which delays the Answer-part to a Question-part, until some other question has been answered. There are also global organization principles that describe how different types of dialogues are initiated and ended.

These regularities in dialogue have led some researchers to propose that coherent utterance exchanges in dialogue can be described by means of conversational rules, much the way coherent sentences are described by syntactic rules. The basic categories of these conversational rules are speech acts, and the general idea is that sequences of speech acts that adhere to the rules are coherent, while the remaining sequences are incoherent. While there are serious theoretical problems with this approach as a general model for human–human conversations (Levinson 1983), it has been successfully applied to the design of human–computer dialogue systems, such as the dialogue games of Power (1979) and Carlson (1983) or the dialogue grammars of Polanyi and Scha (1984) and Jönsson (1993, 1996).

There are two simultaneous information channels in a dialogue: the information channel from the speaker, and the backchannel feedback from the listener. The backchannel feedback indicates attention, feelings and understanding, and its purpose is to support the interaction (Yngve 1970). It is communicated by anything from short vocalizations like “mm” to utterances like “I think I understand”, or by facial expressions and gestures (Goodwin 1981). Jurafsky et al. (1998) presented a computational model that used lexical, prosodic and syntactic cues for automatically distinguishing between the dialogue acts *yes-answer* and three types of backchanneling acts: *continuers*, *incipient speakership* and *agreement*. All of these can be realized by words like “yeah”, “ok”, “mm-hmm”.

Co-operation

Another fairly well agreed upon finding is that most human dialogues are characterized by co-operation (Grice 1975, Allwood 1976). Grice defined the Co-operative 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*”, which is manifested in the maxims of Quantity, Quality, Relation and Manner. Dybkjær et al. (1996) have extended the Gricean maxims to be useful for human–computer dialogues. They added three more aspects that they argued a dialogue system must take into consideration:

Partner asymmetry Provide clear and comprehensible communication of what the system can and cannot do, and how the user has to interact with the system.

Background knowledge The system has to take into account the users' background knowledge and their assumed expectations of the system's knowledge.

Repair and clarification The system should initiate clarification meta-communication if necessary, e.g. if the user input is inconsistent or ambiguous.

Gaasterland et al. (1992) give an overview of the use of Gricean maxims as a starting point for cooperative answering. They describe cooperative techniques for information retrieval that consider both the users' conceptions and their misconceptions.

Grounding and collaboration

Participants in spoken dialogue establish a *common ground* from their past conversations, their immediate surroundings and the current dialogue (Clark and Schaefer 1989, Clark and Brennan 1991). Speakers co-ordinate their use of language with other participants in a language arena in two phases: first an utterance is presented, it is then accepted when the receiver signals that he has received the information. The acceptance is acknowledged by feedback words like "ok", paraphrases of the presented utterance, or by implicit acknowledgments (Traum and Allen 1992). The implicit acknowledgment could be produced by reusing the terms the participant used or by continuing the dialogue in a way that is in accordance with the previous turn.

Collaboration in dialogue is the process where the participants coordinate their action towards a shared goal. This has been formalized in the Shared Plans theory (Grosz and Sidner 1986), where three discourse structures are used: the intentional structure in the form of Shared Plans, the linguistic structure in the form of segments of actions, and the attentional structure in the form of a focus stack. Collagen is a computational model that is based on this theory (Rich and Sidner 1998). Participants in a conversation also collaborate while making references (Clark and Wilkes–Gibbs 1986). A computational model of how users collaborate on referring expressions was proposed by Heeman and Hirst (1995). Traum and Allen (1992) presented a computational model of grounding. They also defined discourse units (DUs) that are built up by single-utterance grounding acts. They extended the speech act theory into the *conversation act theory* that used four discourse levels: turn-taking, grounding, core speech acts and argumentation. This theory was presented in Traum and Hinkelman (1992).

3.2.2. Domain and task analysis

In task-oriented dialogues users talk with the system in order to be able to complete a task. If it is a task that people usually solve by talking to other humans the dialogue designer can use transcripts from human–human interaction as an inspiration when designing the task model. A method of formalizing this work has been proposed by Jönsson and Dahlbäck (2000). They argue that human–human dialogues might be relevant to get an idea of what tasks users would like to solve and how these tasks are related to each other, but that they will not be optimal for building grammars and language models, since people adjust their way of speaking according to the receiver. Their suggestion is to distill the human–human dialogues by re-writing them as the dialogue designer thinks they would have been conducted if they were human–computer dialogues. Their distilling guidelines state that the 'user' utterances should be changed as little as possible, and that the 'system' utterances should be changed in order to match the desired capabilities of the future automated system. These dialogues can give the designer inspiration for the task analysis needed to build the initial domain model and a WOZ system that could be used to collect more realistic data. Another way of getting the information needed for the task analysis from an existing manual version of the service is to interview the humans that perform the task, as well as their customers.

If there is no manual version of the service, other methods have to be used. One commonly used method in system development is scenario analysis. A scenario describes a user situation where an imaginary user interacts with the future system. The scenario can describe different kinds of typical users in various user situations. The scenario describes everything that is happening, what the user does and how the system reacts. These descriptions can be used to specify what kind of modules the system must have. A method that was used in the Olga project (Beskow & McGlashan 1997) at this stage in the scenario analysis was to let the system developers simulate their own modules verbally while stepping through the scenarios. At each step they had to specify what messages their modules would need from other modules in order to perform the tasks needed for them to support the complete system in generating the desired output. A popular type of scenario in object-oriented system development is *use cases* (Jacobson et al. 1992). According to Hulstijn (2000) use cases are useful in the development of simple spoken dialogue since they are easy to use and because they generalize over a set of related scenarios. However, Hulstijn states that use case tools (flow graphs and sequence diagrams) are not optimal for developing mixed initiative dialogue systems.

3.2.3. Empirical user studies

One final source of knowledge on how to build spoken dialogue systems is to perform user studies on people who interact with real or (partly) simulated systems. With such simulations, researchers and system designers achieve many goals: they get data for speech recognizer training, information on domain-specific expressions, utterances and dialogue patterns, turn-taking behavior, and (if the system is multimodal) information on how the users prefer to make use of the various modalities. In addition, data from simulations may point to problems and research issues that would have been difficult to anticipate otherwise.

Rapid prototyping and iterative development are common methods in software engineering, and they can also be useful when developing spoken dialogue systems. These methods have been successfully used for graphical interfaces where the user interacts with a system that is not fully functional, but that gives the users the look and feel of the final system. This is not as easy in spoken dialogue systems since the prototype system has to have some rudimentary speech understanding and the system designer has to anticipate how the users will speak to the prototype. To be able to build the first iteration of the prototype it might be necessary to simulate at least the speech understanding.

The method of simulating the whole or parts of a dialogue systems, in order to be able to collect human–computer interaction, has been called the Wizard-of-OZ (WOZ) method (Richards and Underwood 1984, Guyomard and Siroux 1988, Jönsson and Dahlbäck 1988, Fraser and Gilbert 1991). A crucial part of WOZ simulation is to make the subjects think that they are interacting with a fully automated system. The term comes from the children's novel *The Wonderful Wizard of Oz* (Baum 1900):

> "Presently they heard a solemn Voice, that seemed to come from somewhere near the top of the great dome, and it said, 'I am Oz, the Great and Terrible. Why do you seek me?' They looked again in every part of the room, and then, seeing no one, Dorothy asked, 'Where are you?', 'I am everywhere', answered the Voice, 'but to the eyes of common mortals I am invisible. I will now seat myself upon my throne, that you may converse with me.'
>
> [...]
>
> As the screen fell with a crash they looked that way, and the next moment all of them were filled with wonder. For they saw, standing in just the spot the screen had hidden, a little old man, with a bald head and a wrinkled face, who seemed to be as much surprised as they were. The Tin Woodman, raising his axe, rushed toward the little man and cried out, 'Who are you?', 'I am Oz, the Great and Terrible,' said the little man, in a trembling voice. "But don't strike me--please don't-and I'll do anything you want me to."

There are some requirements that need to be met in order to perform a WOZ simulation. It must be possible for the wizard to perform the intended tasks, the desired system behavior must be specifiable and the whole simulation must be believable (Fraser and Gilbert 1991). An early attempt at dealing with these issues was the ARNE-3 WOZ environment (Dahlbäck et al. 1993). This system had an editor for making database queries and used menus with canned system prompts to ensure fast and consistent answers from the wizard. It is also possible to simulate only one component in the system, like in the Waxholm WOZ experiments where the wizard simulated only the speech recognizer, typing in exactly what the subjects said (Blomberg et al. 1993). The WOZ method for initial data collection has been used in the development of a number of spoken dialogue systems, e.g. Circuit Fix It Shop (Moody 1988), ATIS (Hemphill 1990), SUNDIAL (Peckham 1991), MADCOW (Hirschman 1992), The Philips train timetable system (Oerder and Aust 1994).

If WOZ simulations are used to collect spoken interaction that will be used to train speech understanding components it is important to decide what degree of understanding the wizard should simulate. However, Dahlbäck et al. (1993) did not let their wizards simulate limited understanding because they did not want to decide on what kind of understanding a future system might be capable of. Furthermore, they argued that it is very hard for a wizard to simulate limited understanding in a convincing way. However, Bernsen et al. (1998) argue for giving the wizard simple rules on how to simulate limited understanding, e.g. “do not understand any personal names” or “do not understand negations”. They also state that input filtering can be used to elicit wizard misrecognitions. There have been other attempts at simulating limited understanding capabilities. According to Fraser and Gilbert (1991) text input systems could use filters that introduce insertion and deletion errors, and speech input systems could have a speech recognizer between the subject and the human wizard. Peissner et al. (2001) presented a WOZ system where the human wizard only decided if it was possible for the system to understand the subject’s utterances according to some given restrictions. The system then used this assessment to decide with what probability it should understand the user input correctly.

In multimodal WOZ systems the wizard has to be able to handle the graphical modality as well. In the AdApt WOZ experiment the subjects’ speech and graphical input was sent to a WOZ interface that was run on a seperate computer (Gustafson et al. 2000). Graphical selections were automatically translated to text and inserted into the wizard’s database interface. In the same interface the wizard would insert the constraints that the subject provided in his spoken input. The wizard generated synthesized answers using menus of answer templates. The system could be adjusted to construct verbal answers with deictic references that had to be synchronized with the graphical output. Salber and Coutaz (1993) argue that one could use multiple wizards in multimodal WOZ simulations, one for handling the input, one for handling the task level decisions and one for formulating the answers. They also state that it might be useful to have different wizards for each input modality, but then an additional wizard is needed for modality fusion. However, using multiple wizards in any of these two ways will introduce rather complicated coordination between the wizards, that might be hard to handle.

In the MASK project three cycles of WOZ simulations were run before building the prototype system (Life et al. 1996). They used the WOZ experiments for the dual purpose of prototyping the user interface and for collecting realistic spoken interaction. At later stages of the iterative development of spoken dialogue systems, more advanced WOZ experiments can be conducted. Ammicht et al. (1999) describe a system where the wizard supervises a fully automated dialogue system. The wizard can either control every step of the interaction or let the system work semi-automatically, where the wizard can override the system decisions at different levels when necessary. Thus, the wizard can correct wrong decisions in one module and then let the rest of the system do its job.

3.3. Dialogue taxonomies

A number of taxonomies of dialogues have been presented over recent years. Dahlbäck (1997) distinguishes between seven main dimensions that are relevant when investigating dialogues:

- Modality (spoken or written)
- Kinds of agent (human or computer)
- Interaction (dialogue or monologue)
- Shared context (spatial and temporal)
- Number and types of tasks
- The dialogue-task distance (dialogue and task structures)
- Kinds of shared knowledge

Allen (1997) uses dimensions that describe phenomena a dialogue system must be able to handle:

- Reference resolution - from none to arbitrary anaphora
- Task complexity - from static to hierarchical
- Dialogue management - from none to meta conversation
- Initiative - from fixed single to dynamic mixed

Spoken dialogue systems have so far mostly been designed with an overall goal to carry out a specific task, e.g. ordering tickets. In addition, it would be interesting to consider other goals that the users could have when interacting with a spoken dialogue system. Extending the types of dialogues that could be handled by spoken dialogue systems has consequences for both implementation and evaluation. Most current systems are task-oriented because it makes it easier to build domain models that can be used to pre-define the language models and dialogue rules. Furthermore, having pre-defined tasks makes it easier to evaluate the performance of the dialogue system. The system that can help the users to obtain their overall goal fastest, with the least number of turns might be regarded as the best one.

It would be interesting to extend the goals of spoken dialogue systems, without making it impossible to handle the user interaction that these systems elicit. A first extension could be to remove the overall goal of the dialogues, e.g. buying a train ticket, thus getting explorative dialogues. These dialogues would still have tasks that are solved during the interaction, i.e. giving constraints or receiving information about objects. However, in the case of explorative dialogues, it is harder to compare the number of turns between different user interactions, in order to decide the quality of the system. For how long different users like to browse an information-set varies a lot.

The next step could be to remove the presence of an external task altogether, thereby obtaining interactional dialogues that are not used to achieve anything outside the dialogue itself (Brown and Yule 1983). Correctly designed, such dialogues might still be amenable to processing by a dialogue system, since they are likely to primarily bring up features from the immediate shared context. These context-oriented dialogues would focus on who the dialogue partner is, objects in the shared spatial context and the actual situation of the dialogue partner. The resulting three dialogue categories could be summarized in the following way:

Task-oriented – spoken dialogues that are used to simplify tasks that the users would like to get help with, e.g. control devices and simple computer applications, retrieve or store information in databases, order goods or services, collaborate with the system to do complex tasks. The advantage of task-oriented dialogues is that the turns usually are quite short. Moreover, the users have well-defined goals with their interaction, which make it possible to determine if they have succeeded by measuring task completion rates.

Explorative - spoken dialogues that are used to acquire knowledge about some complex task or browse structured information, e.g. tutoring or educational systems, browsing a large dataset, tourist information systems or asking animated characters in a computer game to perform certain tasks. The users have a goal with their interaction, but it is not easily defined. Rather than finding something in particular it is to explore the possibilities. This makes it hard to determine if and when the users succeed. It is possible to measure query error rates but that does not tell you if the users found what they were looking for. Using an evaluation scheme like Paradise (Walker et al. 1997) might be hard since users might like a system with high query error rate that happened to help them to find something interesting, while they may dislike a system with low query error rate that did not help them to find any interesting items.

Context-oriented – spoken dialogues that are targeted at the dialogue situation, where users engage in small talk with an embodied agent in order to get to know more about the agent's personality, the location where the dialogue takes place and the situation. Bickmore and Cassel (2000) experimented with different degrees of small talk in their REA system to establish a social relation that would increase the users' trust in their animated agent, and to give the users a notion of the agent's capabilities. Context-oriented dialogues could be used by a conversational agent that embodies a real or fictive person in an information kiosk, interactive actors or characters in computer games or museum guides that can engage in dialogues about the exhibition they inhabit. It is quite hard to measure success rate since the user's primary goal is to be entertained. It is possible to measure query error rates, but to build an entertaining system it is not certain that understanding is the only important feature. It might be equally important for the system to be able to act as if it understood or to keep the conversation going trying to understand what the user meant later on. An early example of keeping the dialogue going without deep understanding was Eliza (Weizenbaum 1966). A later example is Julia, a 'chatterbot' who inhabited a text-based MUD (Mauldin 1994, Foner 1997). Julia used word pattern matching and answer templates, often with humorous responses, to create a socially viable persona that users enjoyed chatting with.

The features proposed by Dahlbäck and Allen are of course very important to consider when implementing dialogue systems with any of these three types of goals. Allen et al. (2000) argue that it is necessary to limit spoken dialogue systems to goal-seeking dialogues, *practical dialogues*, which would correspond to the task-oriented dialogues and explorative dialogues above. They suggest that unrestricted natural dialogues are too hard to handle. However, the experiences from the August and Pixie systems indicate that people are rather restricted in context-oriented dialogues as well. The users mostly talked about the agent and the shared spatial context. If a dialogue system would have control of the context it would be possible to build systems where the users' goal is to have entertaining dialogues. These kinds of context-oriented dialogues will be important if spoken dialogues are to be used in computer games. The users will be able to refer multimodally to objects in their shared spatial context. It will be possible to generate context-oriented dialogues since the system will know what is shown in a particular scene and since the personality and traits of the characters will be indicated by their appearance, movements and speech.

Spoken dialogue systems could benefit from having all three types of goals, but the benefit might vary in importance depending on the application type, which is indicated in Table 1. The size of the dots in this table represents an attempt at assessing the importance of the dialogue goals for a range of application types.

The first three application types in this table are typical task-oriented domains that most spoken dialogue systems have been targeted at so far. In these task-oriented domains, explorative dialogues might be useful as a help option, where they could be used to explain the available options in the system. Context-oriented dialogues might be useful if the interface is embodied by a persona with a certain personality (André et al. 1999). The *problem solving* applications built so far have focused on solving a single specific task, e.g. rescuing people (Allen et al. 2001) or mending a circuit board (Smith et al. 1992), but there is of course a need for a browsing feature in these also, e.g. while the users are familiarizing themselves with the domain of the system. *Information browsing* is the obvious application for explorative dialogues. Bickmore and Cassel (2000) argued for the use of small talk in their REA system in order to "grease the wheels of task talk". It would also be natural to use an information browsing system to perform a certain task, e.g. allowing a user of the AdApt system to contact the seller of a specific apartment in order to place a bid on it.

Table 1. *A list of application types, examples of systems, the importance of supporting different types of goals, and the value of an embodied agent.*

Type of application	**Example Systems**	**Task-oriented**	**Explorative**	**Context-oriented**	**Embodied Agent**
Voice-controlled devices	Put-That-There (Bolt 80) VODIS (Westphal & Weibel 99) D'homme (Rayner et al. 01)	●	·	·	·
Transactional Systems	Telia/SRI travel system (Boye et al. 99) CTT-bank (Melin et al. 01) SmartKom (Wahlster et al. 01)	●	·	·	·
Information Retrieval	Sundial (Peckham 1991) Voyager (Zue et al. 91) Waxholm (Blomberg et al. 93) Philips Train timetable (Aust et al. 95) PADIS (Kellner et al. 96) Galaxy (Seneff et al. 98), Arise (Lamel et al. 98) MIMIC (Chu-Carroll 00)	●	·	·	·
Problem Solving	Circuit Fix It shop (Smith et al. 92) Trains (Allen et al. 95), Trips (Allen et al. 01) Larri (Bohus & Rudnicky 02)	●	•	·	·
Information Browsing	REA (Cassell et al. 99) AdApt (Gustafson et al.00) Nokia TV Guide(Ibrahim et al. 01)	•	●	•	•
Tutoring system	PPP persona (Rist et al. 97) Collagen (Rich et al. 01) Steve goes to Bosnia (Traum & Rickel 01)	•	●	•	●
Educational system	Herman the Bug (Lester & Stone 97) Baldie (Cole et al. 99) Cosmo (Lester et al. 99) Steve (Johnson et al. 00)	•	●	•	●
Persona/Guide at Exhibition/Museum	August (Gustafson & Bell 00) I SEE (Oviatt 00) Mack (Cassell et al. 02) Pixie (Gustafson & Sjölander 02)	·	●	●	●
Entertainment, Computer Game	Seaman (Sega 00) Hey, you, Pikachu! (Nintendo 99) NICE (www.niceproject.com 02)	·	●	●	●

Both tutoring and educational dialogues are primarily explorative, but they have the overall goal of teaching somebody something, maybe to improve a student's performance in a subsequent written assignment. An embodied character with a personality that can engage in context-oriented dialogues might make the learning experience more engaging, thus improving the learning effect (Dehn and van Mulken 2000).

The last two application types are the most context-oriented. Initially the users will try to get to know the animated characters. Then they will explore and use the agent's capabilities, e.g. telling it to do things in a game or asking it questions about an exhibition. An embodied character in an exhibition could also be used to perform tasks, such as ordering tickets for certain events, and the players of dialogue games might have the overall goal of solving the game.

This chapter has described the parts needed to build multimodal spoken dialogue systems, and has outlined some guidelines on how to build them. The importance of studying the way humans interact with other humans as well as with computers has been stressed. Lastly, a taxonomy of dialogue goals was presented and discussed. The next chapter will describe five spoken dialogue systems that have been developed as part of the current thesis.

Chapter 4

Five dialogue systems

The work described in this thesis was carried out at the Department of Speech Music and Hearing (TMH) at KTH between 1992 and 2000, and at Telia Research between 2000 and 2002. TMH has a long-standing tradition as one of the world's leading speech research departments. The pioneering work of Fant, Lindblom, Öhman and others in the fields of speech production and perception (Jakobson, Fant and Halle 1952, Fant 1960, Lindblom 1963, Öhman 1966), was later applied in work on speech synthesis and recognition (Fant 1953, Liljencrants 1967, Carlson and Granström 1976, Blomberg and Elenius 1978). Since 1992, the department has also worked in the dialogue field, building spoken dialogue systems. Many of these systems have used an animated talking head as a dialogue partner for its users. The spoken dialogue systems that will be described in this thesis are shown in Figure 7. Their chronology is indicated by the timeline: The Waxholm system was developed from 1992 to 1995, Gulan from 1997 to 1998, August in 1998, AdApt from 2000 to 2002 and finally the Pixie system that was developed at Telia Research in 2002.

This chapter introduces these systems briefly and gives an overview of some of the features of the systems that might have influenced the dialogue corpora collected. All systems have used spoken Swedish as input and output, and all dialogue examples below have been translated into English.

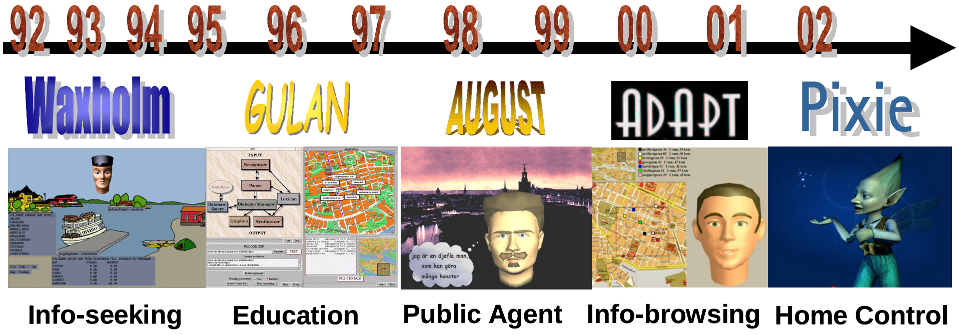


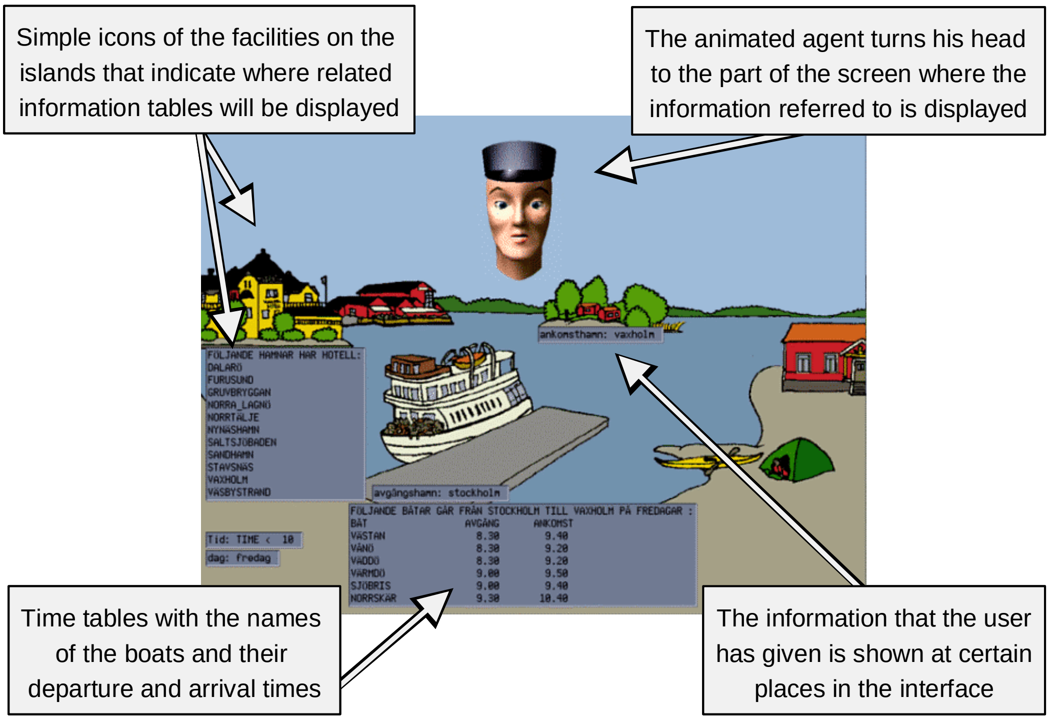
Figure 7. *The spoken dialogue systems presented in this thesis.*

4.1. Overview

This section will give a short overview of the five dialogue systems, as well as some example dialogues that will give an idea of their functionalities.

4.1.1. Waxholm

Waxholm was a spoken dialogue system for retrieving information about the ferryboat services in the Stockholm archipelago (Blomberg et al. 1993). The system also had some information about facilities like hotels and restaurants on the islands. The system could be used to find information, not to carry out the actual bookings or provide further tourist information. The system featured a graphical interface with an animated talking head and a picture that visualized the system's domain (Bertenstam et al. 1995). Textual information was presented by placing tables by the icons depicting the corresponding facilities, as in Figure 8, where the table with available hotels is below the picture of the hotel, while the timetable is shown below the boat. Information provided by the user was also displayed at different places, the recognized destination was shown on the island and the recognized departure on the jetty.



Simple icons of the facilities on the islands that indicate where related information tables will be displayed

The animated agent turns his head to the part of the screen where the information referred to is displayed

FÖLJANDE HAMNAR HAR HOTELL:
DALARÖ
FURUSUND
GRUVBRYGGAN
NORRA_LAGNÖ
NORRTÄLJE
NYNÄSHAMN
SALTSJÖBADEN
SANDHAMN
STAVSNÄS
VAXHOLM
VÄSBYSTRAND

ankomsthamn: vaxholm

avgångshamn: stockholm

Tid: TIME < 10

dag: fredag

FÖLJANDE BÅTAR GÅR FRÅN STOCKHOLM TILL VAXHOLM PÅ FREDAGAR :

BÅT	AVGÅNG	ANKOMST
VÄSTAN	8.30	9.40
VÅNÖ	8.30	9.20
VADDÖ	8.30	9.20
VÄRMDÖ	9.00	9.50
SJÖBRIS	9.00	9.40
NORRSKÄR	9.30	10.40

Time tables with the names of the boats and their departure and arrival times

The information that the user has given is shown at certain places in the interface

Figure 8. *The Waxholm user interface.*

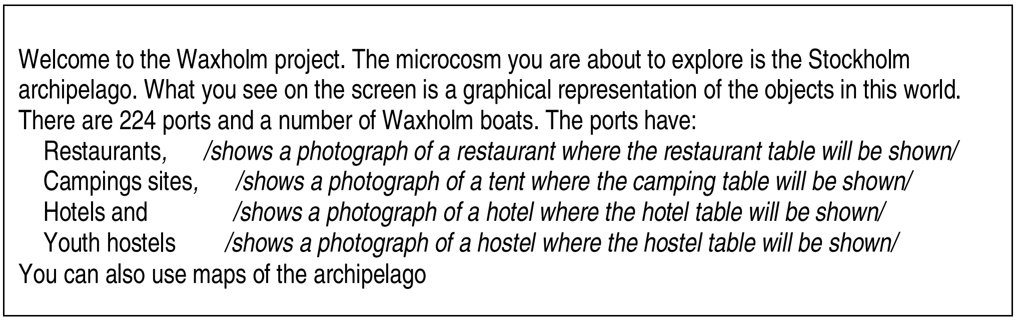
The Waxholm project was initiated in 1992 as a research effort for building spoken dialogue systems. In this project, new dialogue management and parsing modules were developed and combined with TMH’s existing speech synthesis and recognition. The goal was to acquire knowledge on how to develop the natural language modules and the other system modules needed to build spoken dialogue systems (Carlson 1996). Another important purpose was to collect spoken dialogue data. The fully automated Waxholm system has not been used in any extensive users studies. These were instead carried out during the iterative development of the system. The WOZ version was used to collect dialogues from 66 subjects. Before interacting with the system, the subjects of the WOZ experiment were given written scenarios, but these turned out to influence the subjects. Some of the subjects reused large parts of the written scenarios when they approached the system. An example of this can be seen in Figure 9.

> Scenario 2:
> You would like to take a trip **out into the archipelago during the weekend**. Because you’d like to have a comfortable stay, you want **to stay at a hotel.** You quit work at 3:00 pm on Friday and start work again at 10:00 on Monday. Find out where you can stay and when you can travel.
>
> The user’s first utterance:
> ”I want to go **out into the archipelago during weekend** EEH and **stay at a hotel**, when, where is it possible EEH to stay”

Figure 9. *Example of priming from the written scenarios in Waxholm.*

The users were also influenced by the wording of the verbal output of the system. This phenomenon was further investigated in the study that is presented in *Paper II* of this thesis. In this study users had to answer questions about their holiday plans. The system used one of two verbs in its questions. The subjects reused the verb in their answers in 51% of the cases while they used another verb with the same or almost the same meaning in only 4% of the cases (in the rest of the cases they did not use any verb at all in the answer or did not even answer the question).

Another problem was that the users often did not understand what the system was looking for. The system did not use verbal confirmation in each turn since it would make the dialogues slow and unnatural. To overcome the priming problem and to be able to give the users feedback on what the system had understood so far, a graphical representation of the system domain was added (Bertenstam et al. 1995). The domain can be viewed as a microcosm consisting of harbours with facilities and boats that can take you between them, see Figure 8. Instead of the written scenario the animated talking head gave a verbal introduction to the system while the system highlighted the relevant parts of the graphical interface, as can bee seen in Figure 10.



Welcome to the Waxholm project. The microcosm you are about to explore is the Stockholm archipelago. What you see on the screen is a graphical representation of the objects in this world. There are 224 ports and a number of Waxholm boats. The ports have:

Restaurants, */shows a photograph of a restaurant where the restaurant table will be shown/*
Campings sites, */shows a photograph of a tent where the camping table will be shown/*
Hotels and */shows a photograph of a hotel where the hotel table will be shown/*
Youth hostels */shows a photograph of a hostel where the hostel table will be shown/*

You can also use maps of the archipelago

Figure 10. *The verbal introduction that the talking head gave while the pictorial introduction was shown in the GUI.*

The idea of the pictorial introduction was to give the users a hint of what kinds of things they could ask about and also to remind them of this later on in the dialogue. Another purpose was to be able to continuously feedback the information that the system had obtained from the processing of the users' utterances, such as place of departure, day of travelling and so on. The interface was also meant to give a graphical view of the knowledge the subjects had secured so far, in the form of listings of hotels, etc.

The nature of the domain, with boats cruising in the Stockholm archipelago - sometimes returning to the same harbour on the same trip, made the backend system a bit complicated (Gustafson 1992). Another feature of the domain was the hierarchal structure, were harbours were located on islands. As an example of problems resulting from the hierarchal structure, *Stockholm* was regarded as an island by the system, but probably as a harbour by the users. This solution was chosen because it could not be expected of the users to know which of the three harbours in Stockholm they had to go from in order to be able to get to their destination. Dialogue 1 below shows an example of when this led to confusing system output. It is a user interacting with an early version of the WOZ system, where the notion of return-trip had not been implemented yet. Since all harbours in Stockholm were considered when the users asked for Stockholm, the system in this example tries to find a trip between any of its harbours. However, the subject identifies this problem and corrects it at the same time as he specifies on what day he wants to return.

SUBJECT	SYSTEM (Waxholm)
Er I have a question	I don’t understand
Where is Lådna?	Lådna is located on this map
Er how long does it take to go there	From where do you want to depart?
From Stockholm	Where do you want to go?
To Lådna	Which weekday do you want to travel?
On Friday	There are many boats that go from Stockholm to Lådna on a Friday, what time do you want to go?
Well... in the afternoon	This is a timetable of the boats that go from Stockholm to Lådna on a Friday after 12 o’clock am and before 6 o’clock pm .
Are there no boats later at night?	This is a timetable of the boats that go from Stockholm to Lådna on a Friday after 5 o’clock and before 12 o’clock.
If I want to go back to Stockholm later on?	There are no boats that go from Stockholm to Stockholm on a Friday after 5 o’clock and before 12 o’clock.
From Lådna on the Saturday?	This is a timetable of the boats that go from Lådna to Stockholm on a Saturday after 5 o’clock and before 12 o’clock.

Dialogue 1. *A dialogue example from the Waxholm WOZ corpus. The subjects were given the following written scenario: “You are visiting friends who have a summer cottage on Lådna. Your task is to locate Lådna and then find out how to get there. Please start by pushing the button.”*

4.1.2. Gulan

The Gulan educational system was a modular dialogue system which students could reconfigure interactively whilse running it. The students got a simple dialogue system where they could search the web-based version of the Swedish Yellow Pages. The system presented information both verbally (with an optional talking head) and graphically on a map and in a text table. The student could inspect and change the different modules of the spoken dialogue system. The purpose of the lab assignment was to stimulate the students to think about the possibilities, limitations and some practical problems of task-oriented spoken dialogue systems in the information retrieval domain. The system was developed at KTH (Sjölander and Gustafson 1997). A new dialogue manager for Gulan was developed by NLPLAB at Linköping University in a joint project (Gustafson et al. 1998). Gulan has been used in a number of speech technology courses at five different universities in Sweden. The system has also been demonstrated live at a number of workshops.

The system overview window is shown to the left in Figure 11. The students could access the different dialogue modules by pressing the buttons at the bottom of the screen.

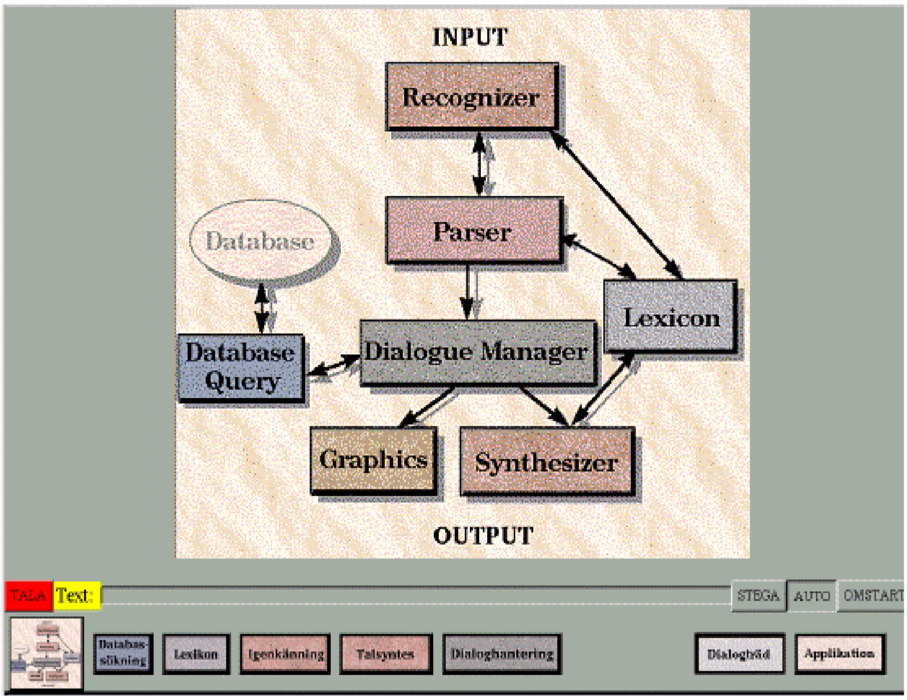
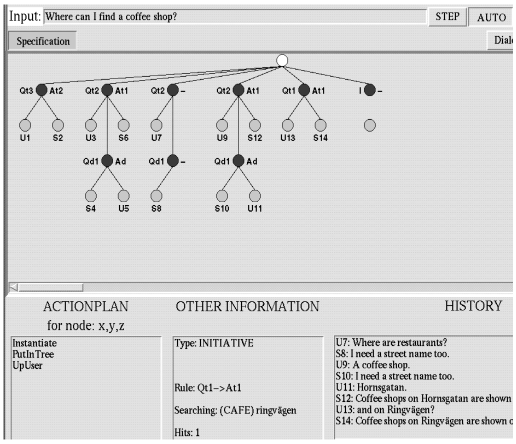


Figure 11. *The Gulan interface, where the overview window is shown.*

Gulan has been used in lab assignments where students were given a version of the system which had some basic functionality that they could extend. They were told to first try the system for a while, then later to extend the lexicon with more words, add new items, e.g. book stores, from the yellow pages database and finally add new dialogue management rules. The lexicon module was the most central one, where the students would make many of their changes and additions. It was used by the recognizer, the keyword spotting module, the dialogue manager and the speech synthesizer. The lexicon had five fields: orthography, transcription(s), semantic class, semantic sub-class and a class/instance tag that was used by the dialogue manager.

In the recognition module the students could generate a new recognition lexicon that they could use on a previously recorded utterance or interactively by pressing the push-to-talk button and producing a new utterance. Some of the pruning parameters could be changed during runtime. The students could generate a ten-best list and get information about the CPU-time it took to generate it. They could try different settings to see for themselves how different pruning parameters influence the speed and quality of the speech recognition. They could also visualize the speech file with f0-contours and spectrograms, which made it possible see how different ways of speaking and background noise influence the recognition rates.

The dialogue manager module, which is a simplified version of the LINLIN dialogue manager (Jönsson 1997), was implemented by Patrik Elmberg (Gustafson et al. 1998). The students could change the dialogue rules and the focus handling strategy as well as inspect the dialogue history tree by going to the dialogue history window shown in Figure 12.



Input: Where can I find a coffee shop?

Instantiate
PutInTree
UpUser

Type: INITIATIVE

Rule: Qt1–>At1

Searching: (CAFE) ringvägen

Hits: 1

U7: Where are restaurants?
S8: I need a street name too.
U9: A coffee shop.
S10: I need a street name too.
U11: Hornsgatan.
S12: Coffee shops on Hornsgatan are shown
U13: and on Ringvägen?
S14: Coffee shops on Ringvägen are shown c

Figure 12. *The dialogue history window of the Gulan system.*

Gulan has been used in courses at five different departments since 1997. In 1998 it was used by a total number of 150 students. The learning effect of the dialogue component of Gulan has been evaluated by Qvarfordt and Jönsson (1999).

4.1.3. August

The August system was a conversational spoken dialogue system featuring an animated agent called August, whose persona was inspired by August Strindberg, the famous Swedish 19th century author (Gustafson et al. 1999). The August project was initiated as a way to promote speech technology and KTH in connection with Stockholm being the Cultural Capital of Europe in 1998. The spoken dialogue system as well as the animated character were developed during the first half of 1998 and the system was available for the general public at the Culture Center in Stockholm, daily from August 1998 to March 1999, see Figure 13.

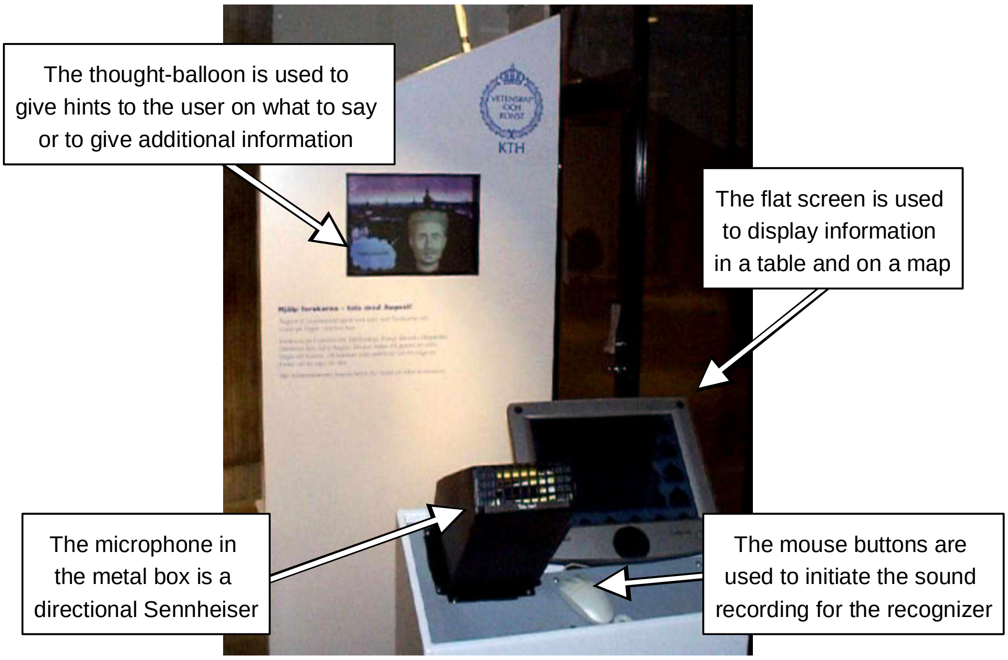

The thought-balloon is used to give hints to the user on what to say or to give additional information
KTH
The flat screen is used to display information in a table and on a map
The microphone in the metal box is a directional Sennheiser
The mouse buttons are used to initiate the sound recording for the recognizer

Figure 13. *The August display case at the Cultural Center in Stockholm.*

The research purpose of the project was to expose speech technology to the general public, and in this way get practical experience from moving a research system outside the lab environment, and at the same time collect data on how people might interact with animated agents. August could answer questions covering a number of topics, for example giving the location of restaurants in Stockholm, sharing facts about the author August Strindberg or exchange social utterances. The dialogues can be considered as quite shallow since the system primarily answered questions and only occasionally initiated one-level clarification sub-dialogues. This meant that the dialogues were user driven, which of course influenced the dialogue data collected.

August was a spoken dialogue system with multiple domains. The first issue that had to be handled was how the system should communicate which domains it could handle, without explicitly asking the users to ask certain questions. To make it possible to give hints on topics of conversation, a thought balloon was added. If the user asked August something that he did not understand August would state that he did not understand, while at the same time indicate that he was 'thinking' by displaying *Why don't they ask me about Strindberg?* as text in the thought balloon. The users would also ask August what they could talk about, see Dialogue 2. As can be seen August would in these cases tell the users about the things he could talk about.

USER (Child)	SYSTEM (August)
What do you know?	I can tell you almost anything.
What can I ask you?	Do I seem confused?
Yes!	I did not understand that!
What is your name?	I did not understand that!
What is the name of the Swedish king	I can't tell you about such things!
What can I ask about then?	I know where to find restaurants in Stockholm
How many restaurants are there in Stockholm?	I did not understand that!
What restaurants are there on Hornsgatan?	The information you asked for is shown on this map /shows a map on the second screen where the found restaurants are displayed/

Dialogue 2. *An example from the August corpus.*

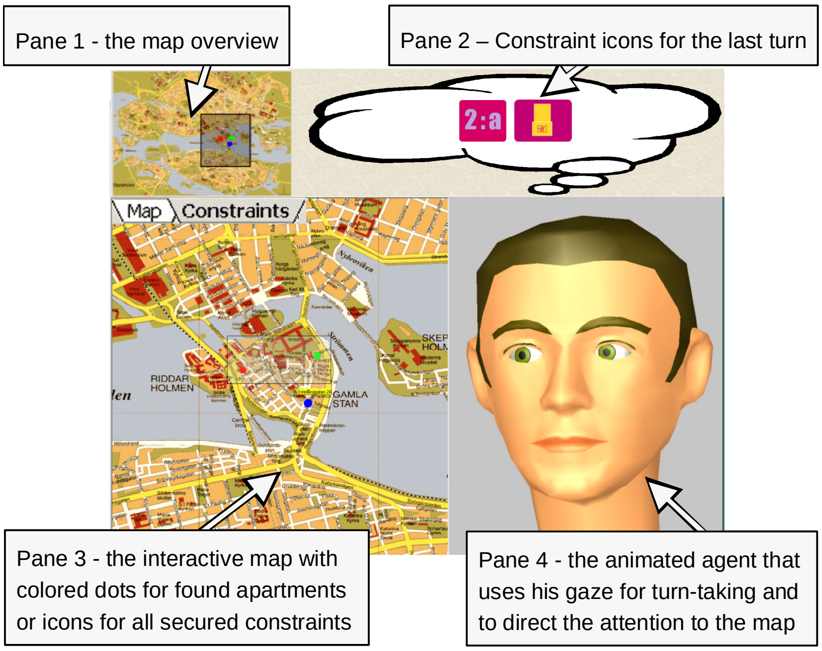
The initial idea was that the different domains should have different kinds of dialogue managers, but due to limited time and resources all different domains used the same dialogue manager, but with different domain specific knowledge bases. However, this separation facilitated structuring the knowledge in each domain and it also made it easier to add or delete domains. The dialogue manager was quite rudimentary since it only handled question–answer dialogues with a single level of clarification questions, e.g. *Where can I find restaurants? – You have to specify a street for me to be able to help you.* The system used machine learning to do domain prediction, semantic analysis and to filter out implausible hypotheses from the speech recognizer (Lindberg and Gustafson 2000). This method was used to facilitate rapid prototyping and simplify an extension of the system according to the collected user interaction.

One of the lessons learned from the August project was the difficulty of handling speech input in public spaces, with tough acoustic conditions. The August system was exhibited in a public space with a stone floor, glass walls, and background noise from other equipment and visitors constantly passing by. Because of the acoustic conditions it was necessary to use a *push-to-talk* solution for speech input, instead of using speech detection. The simplest microphone solution would have been to use a headset, but this was considered too vulnerable in a publicly available system. Instead, a number of ways to mount a microphone out of reach from the users were considered (Gustafson et al. 1999). An initial idea was to use an acoustical lens in the form of a large balloon, filled with CO_2 .The speaker would then stand in front of the balloon and the microphone would be placed at the focal point at the other side. This did not work well because the sphere had poorly defined focal points and insufficient effect at low frequencies. A second trial was to build a large segment of an ellipsoid reflector, where the focal points were located at the speaker position and 1 m above, respectively. Again, the basic problem was that the size of this reflector was too small to have an appreciable effect below about 1 kHz. Getting sharp enough focussing would require a bigger reflector than was possible to set up. The solution finally selected was to use a directional microphone, secured in a metal grid box, into which the speaker could talk at short distance. The box introduced some deterioration of the sound but this did not affect the recognition significantly.

August used facial gestures for a number of purposes in the dialogue (Lundeberg and Beskow 1999). He would typically raise his eyebrows early in the sentence followed by a small nod, and he would mark focal words and stressed syllables with eyebrow movements. To enhance the perceived reactivity of the system, a set of listening gestures and thinking gestures was used. When the user pressed the push-to-talk button, the agent immediately started a randomly selected listening gesture, e.g. raising the eyebrows. At the release of the push-to-talk button, the agent changed to a randomly selected thinking gesture, e.g. looking away from the user. In order to make the synthetic face appear less artificial, and to make the agent appear to be aware of the user's actions the agent changed the direction of the head and eyes according to the detected movements of an approaching user. This was accomplished by using a desktop video camera together with image analysis software (Öhman 1999).

4.1.4. AdApt

The AdApt system was a multimodal spoken dialogue system in an information-browsing domain. The users could get an overview of available apartments in Stockholm by means of interaction with a virtual real-estate agent (Gustafson et al. 2000). The practical goal of the project was to develop a conversational dialogue system where the interface would be as intuitive as possible and where the system would be multimodal both in the input and the output channels. The users could for example refer to apartments verbally or graphically by selecting them on an interactive map. One of the research goals of the project was to study how people interact multimodally, and how the design of the system output would influence the users' input behavior. Another goal was to see what the system requirements are on a spoken dialogue system that can handle multimodal and conversational input. The system was also developed to make it possible to experiment with different ways of using an animated character in the dialogue, e.g. to handle turn-taking using appropriate gazing and head movements. The system was developed in two phases. A WOZ version of the system was developed in the spring of 2000 and multimodal dialogue data from 32 subjects was collected. The data was then analyzed and used in the development of the fully automated version of the system that was completed in March 2001. The system has been further developed in 2002 and the final system is shown in Figure 14.



Pane 1 - the map overview

Pane 2 – Constraint icons for the last turn

Pane 3 - the interactive map with colored dots for found apartments or icons for all secured constraints

Pane 4 - the animated agent that uses his gaze for turn-taking and to direct the attention to the map

Figure 14. *The AdApt user interface.*

The users of the AdApt system could interact with the animated agent Urban to browse apartments available in Stockholm. To simplify the browsing the system used reference handling that allowed the users to keep either the object or the question from the previous turn when asking the system a new question. This made it possible for the users to ask a number of questions about a certain apartment without having to explicitly refer to it in each turn, see for example Dialogue 3.

USER	SYSTEM (AdApt)
Available one room apartments on södermalm	How much are you willing to pay?
1.200.000	Is there anything special you would like your apartment to have
No	There are 7 such apartments, these are shown on the map
Bellmansgatan 32	NON-CLOSING GESTURE
The apart-(cut-off)	NON-CLOSING GESTURE
How much does the apartment on Bellmansgatan 32 cost?	The black apartment costs 890000 crowns
How many square meters?	The black apartment is 37 square meters
Which floor?	The black apartment is on the fourth floor
How many rooms?	The black apartment has one room
How many square meters is the yellow apartment?	The yellow apartment is 40 square meters
How much does it cost?	The yellow apartment costs 895000 crowns
Which floor is it on?	The yellow apartment is on the first floor
How many rooms?	The yellow apartment has one room
How much does the apartment on Bastugatan 57 cost?	The green apartment costs 950000 crowns
How many square meters is the blue apartment?	The blue apartment is 32 square meters
How much does the red apartment cost?	The red apartment costs 750000 crowns
How many square meters?	The red apartment is 44 square meters

Dialogue 3. *An example with a user that interacts with the fully automated AdApt system to get information about some apartments.*

The users could also keep the topic from the previous question. If the system had provided information about a specific apartment in the previous turn, a reference to an apartment was then considered as an elliptic query to get the same information for the apartment newly referred to, e.g. *How much does the blue one cost? –Two million! –The red one? – One million!.* However, in cases where the system had introduced a number of apartments in the previous turn the system would not regard a reference as a complete turn. The AdApt system used a method described in *Paper VIII* to decide whether the user's utterance was complete in the current discourse. This method used syntactic cues as well as information from the dialogue manager on what types of utterances should be considered complete in this dialogue context.

The AdApt system used facial gestures for turn-taking, some of which are shown in Figure 15. If an utterance was considered incomplete the animated agent would display the continued attention gesture. If the user had not continued within four seconds the system would timeout and try to interpret the utterance fragment anyway, probably asking a clarification question. In cases where the input was considered complete the agent would make the turn-taking gesture while generating the answer, and then look towards the user again while speaking. After finishing speaking the agent would indicate that his attention was on what the user might say next.

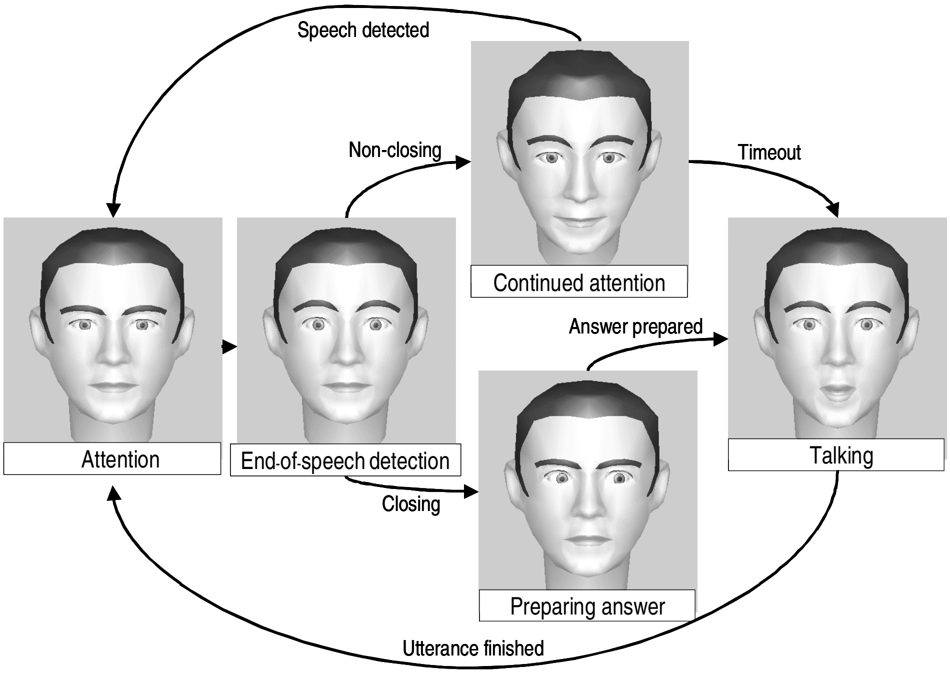


Figure 15. *Turn handling gestures in the AdApt system.*

Another way of indicating that the system had understood a turn was to show icons that represented the constraints in the users last utterances that the system had been able to understand. This was chosen instead of verbal feedback to make the interaction faster and to make it easier for the user to detect and correct misunderstandings. To make browsing the information easier, the user also got an overview of all constraints the system had understood so far in the dialogue, thus making it possible to remove or change previously given constraints. This constraint visualization is described in *Paper IX*.

The AdApt system referred to the found apartments by the color of the dot that indicated their position on the interactive map. A study was performed where the users also interacted with a system that used deictic references, e.g. highlighting an icon while saying *This one costs two million*. This WOZ experiment showed that it was easier to prime the users to refer to apartments by their colors than to influence them to use deictic expressions (Bell et al. 2000). Since it is easier to recognize seven colors than seven hundred street names the initial version of the fully automated AdApt system used colors when referring to apartments, see Dialogue 4.

USER (fp08)	**SYSTEM (Urban)**
I am looking for an apartment in the center of Stockholm	How many rooms would you like to have?
I would like to have two rooms and a kitchen	How much can the apartment cost?
Four million	Where should the apartment be located?
In Vasastan	Is there anything special you would like your apartment to have
Parquet floor	There are 7 such apartments, shown on the map
I am interested in **the apartment at Roslagsgatan**	NON-CLOSING GESTURE
I want information about **the apartment on Rödabergsgatan**	NON-CLOSING GESTURE
How much does **the apartment on Frejgatan** cost	The pink apartment costs 1695000 crowns
How much does **the 'rose' apartment** cost	The pink apartment costs 1695000 crowns
How much does **the black apartment** cost	The black apartment costs 1695000 crowns
How much does **the white apartment** cost	The white apartment costs 1195000 crowns
Does **it** have an open stove?	Yes the apartment has an open stove
How old is **the house**?	The white apartment was built in 1884

Dialogue 4. *A user that is primed to use colors while referring.*

This dialogue is taken from a study where 26 users interacted with the first version of the fully automated AdApt system (Edlund et al. 2002). This corpus shows the same priming effect that was found in the earlier WOZ study. The user in this example realized that the system had difficulties in recognizing street names, and immediately picks up the way of referring that the system uses. The verbal priming can be shown more clearly by looking at all referring expressions produced by all users. Figure 16 shows how all users referred to apartments in their first forty references. In the first ten references the users did not have a preferred way of referencing that they used more often. The share of elliptic questions like *What is the price?* or questions with pronouns like *Does it have a bathtub?* does not seem to change during the course of the dialogues. The system referred to the apartments by color in all its answers, which might explain why the share of color references increases at the expense of the references using the address or other constructions.

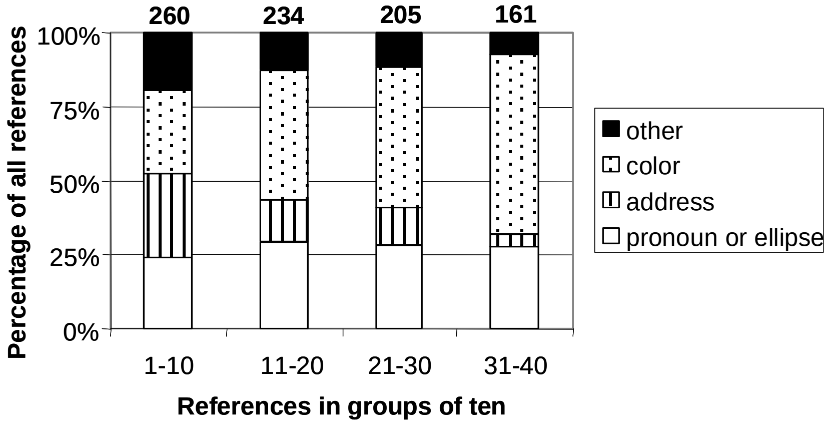
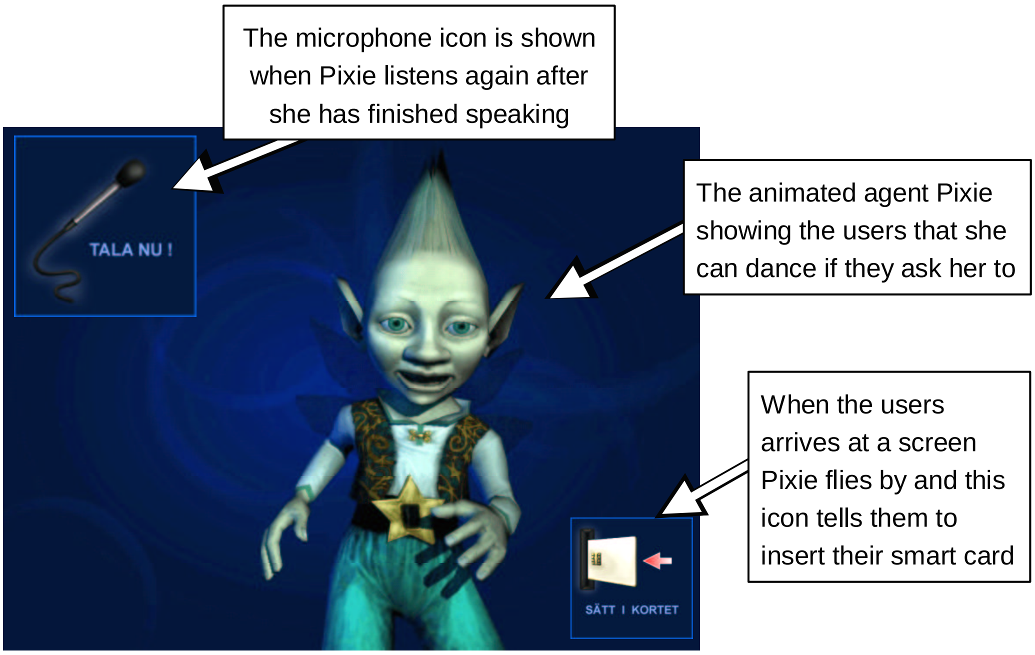


Figure 16. *The priming effect from the system's use of color references. Each bar represents the references from all users. The total numbers of all references in each group of ten are given at the top of the picture.*

Another priming effect was found when analyzing the very first utterances from the subjects in the AdApt WOZ experiments and the AdApt user test, respectively. In the WOZ study the subjects were given pictorial scenarios with about four constraints each that they should specify, where the numerical constrains could be an interval like 2–3 rooms or built between the years 1890–1920. The users in the test of the initial version of the fully automated system on the other hand were only asked to look for an apartment. The first utterance from the subjects in the WOZ corpus was on average 10,4 words long, while it was only on average 5,7 words long for the users in the system corpus. They had exactly the same interface and the same overall task, the only difference being that the WOZ subjects were given some suggestions on what to look for in the pictorial scenarios. Both groups specified about one constraint in their first utterance. However, the WOZ subjects were primed by their pictorial scenarios to provide multiple values for the constraints. They often said things like *Hi I'm interested in a two or three room apartment on Södermalm or Gamla Stan .. it should preferably be built before 1920 maybe as early as 1790. Somewhere in between.* This behavior was not found at all in the dialogues collected with the initial AdApt system, where the users were not given any suggestions on what to look for. These users would instead provide one constraint value per slot and ask about the found apartments, and then change the constraint to get new sets of apartments to investigate.

4.1.5. Pixie

Pixie is an animated agent with whom visitors can talk at the Telecommunication museum in Stockholm. The system is placed in a permanent exhibition called “Tänk Om” (“What If”) that opened in January 2002. The exhibition consists of a full-scale future apartment, aimed at showing what such a home might look like in the year 2010. Among other things the visitors can interact with Pixie, who is shown in Figure 17. Pixie is there to show the potential of embodied speech interfaces, and how they could be used in the future to control devices in the home and to retrieve information. The visitors are asked to either assist Pixie in performing certain tasks in the apartment or to ask Pixie general questions about herself or the exhibition. The users can also ask Pixie to change the lighting in the apartment.



The microphone icon is shown when Pixie listens again after she has finished speaking

The animated agent Pixie showing the users that she can dance if they ask her to

When the users arrives at a screen Pixie flies by and this icon tells them to insert their smart card

Figure 17. *The Pixie graphical interface.*

Visitors enter the exhibition in groups of up to 25. There are twelve computer screens that have been built into walls and tables in the apartment, through which Pixie flies as long as nobody is interacting with her. The users have to register before entering the exhibition, providing some basic information about themselves, such as gender and age. The users get the attention of Pixie by inserting a smart card into a slot besides the screen. This makes it possible for the distributed system to load information from the central server about the current user, such as what they have done at the previous screens along with some personal information. The information about the age of the visitor is used to make it possible to transform children's voices, with regard to their frequency distribution, before sending them to a black-box speech recognizer, thus improving the performance (Gustafson and Sjölander 2002).

The users interact with Pixie in different parts of the apartment, see Figure 18. As can be seen these terminals sometimes are very close to each other, which of course makes the acoustic environment problematic.

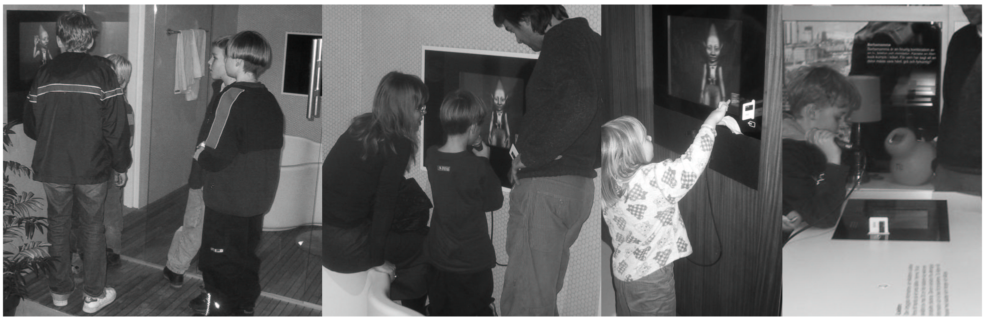


Figure 18. *Some users interacting with Pixie at the different terminals.*

Each terminal has its own list of possible dialogues that it can initiate. Some of the dialogues are only possible at certain places, while others can be initiated from any terminal. The bedroom is for example the only place where the users can change between different lighting settings. Pixie will initiate the following system directed dialogue the first time any of the users insert their card in the bedroom, see Dialogue 5.

USER	SYSTEM (Pixie)
/user inserts card/	Do you want to help me to prepare for the party tonight, maybe by changing the lighting to something more entertaining?
Disco lighting!	Here are some exciting variants: cozy lighting, night lighting, reading light, the default lighting of the Ström family or no lighting.
The default lighting of the Ström family!	The default lighting has been activated, I know how the Ström family want things. Do you want another lighting?
Cozy lighting	The cozy lighting is activated. Do you want another lighting?
Turn on k9 (the robot dog)	Now I don't understand a thing!
Can you do anything else except the lighting?	The lighting is turned off!
Sing me a song!	

Dialogue 5. *An example from the Lighting scenario in the Pixie system.*

There are a number of small task-oriented dialogues that can be initiated at any terminal. In these dialogues Pixie will describe her problem to the user and ask them for advice on what to do. The dialogues are system directed and look something like the one in Dialogue 6.

USER	SYSTEM (Pixie)
/User inserts card/	Do you want to help Peter in his preparations for dinner tonight?
Yes	I know Nancy's absolute favorite dishes. Do you want me to tell you?
/Silence for 5 seconds/ hi	Please help out, Peter would really appreciate it!
Okay, but only because you are so cute!	Here are Nancy's favorite dishes: Sushi and Bangalian stew. Help Peter to choose one of these dishes.
Sushi	Peter is allergic to shellfish. Should I order one of the sushis without shellfish?
Yes do that, its quite alright	Sorry?
/7 seconds of silence/ Hello	I did not quite get that, I'll repeat the question: Should I order …
Yes!	Okay, I'll send the order and then help Peter to prepare dinner. Thank you for your assistance.

Dialogue 6. *An example from the task of helping Pixie to prepare a dinner.*

After one or two such system directed dialogues Pixie will invite the users to ask questions about Pixie herself or the exhibition she inhabits. This results in system directed dialogues where the users can either ask questions or turn on and off devices like lamps. A typical dialogue might look something like Dialogue 7.

USER	SYSTEM (Pixie)
Do you know how to cook?	You look like the tastiest thing in here. Now I 'm coming to gobble you up, hee hee!
What do you know about me?	You are the best, right?
How old am I?	Sorry, I did not quite catch that!
How old am I	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!
What is my name?	You are the best, right?
Do you have a family?	I was created by a bunch of researchers at Telia Research, but my family is of course the Ström family…

Dialogue 7. *An example of the user directed dialogues about Pixie and the 'Tänk Om'-exhibition.*

Initially, the turn handling was a problematic issue in the Pixie system. Due to time lags in network communication, Pixie did not listen until she had finished talking, sometimes with a lag of a little less than a second. This sometimes resulted in the initial part of the user utterances being cut-off. To solve this a listening icon with a microphone was added to indicate when the users could talk. This reduced the average number of too early utterances from about 12% to about 3%.

Wooffitt and MacDermid (1995) describe a simulated spoken dialogue system that used a turn-giving beep to signal that the system gave over the turn to the user. They found that the users talked before the beep more often when the system had failed to handle the previous utterance. The same tendency is found in the Pixie system, where the turn-giving icon is used instead of a beep. As was concluded earlier, the turn-taking icon reduces the number of too early utterances considerably. In problematic discourses users talk too early more frequently. Figure 19 shows that even with the icon people will start talking too early twice as often if the previous system turn was problematic, than if it was correctly handled by the system. A problematic turn would be if Pixie gave the wrong answer or stated that she did not understand the previous user utterance due to a rejection by the speech regognizer. Wooffitt and MacDermid argue that this phenomenon could be explained by the structural properties of repair in conversation (Schegloff et al. 1977). Repairs in human–human dialogues are mostly initiated within the same turn as the problem or immediately at the next transition relevance point (Schegloff et al. 1974). The other speaker will try to position the repair as closely as possible to the error. In cases where the spoken dialogue system fails to correctly handle a turn the users will initiate error recovery as soon as possible, overriding any instructions to wait for the system's turn-giving signal.

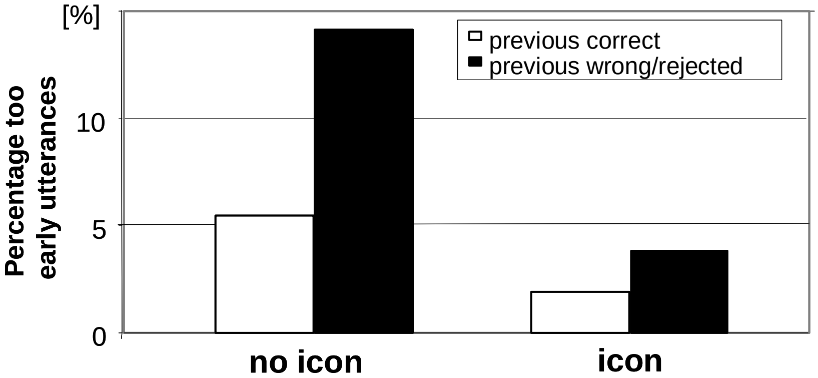


Figure 19. *The percentage of too early utterances depending on whether or not the previous utterance was handled correctly.*

4.2. System requirements

In the dialogue systems described in this chapter, slightly different architectures were used, depending on the requirements of each new system. It has proven quite useful to change both domain and setting of the dialogue systems, since it has forced the system architecture to be reconsidered. Real-time spoken dialogue systems require fast computers with a lot of memory[1]. This means that the design choices are dependent on the capabilities of the computers that are available at the time of development. It is not until the last couple years that it has been possible to build advanced real-time spoken dialogue systems.

The programming languages available are another influential factor. When Waxholm was developed in 1992 all modules were written in C. The communication between the modules was achieved by developing a C-library for message handling. Since 1996 the scripting language Tcl/Tk, which simplifies the development of graphical interfaces, has also been used at TMH. Its embedded nature makes it easy to add speech technology specific extensions, e.g. for audiovisual speech synthesis (Beskow 1997), speech recognition (Ström 1997) and audio handling (Sjölander 1997). Another useful feature is its plug-in capabilities for web browsers, which made it easier to introduce Gulan into courses at other universities (Carlson et al. 1998).

The development of 3D graphic cards has also been dramatic over the last ten years. When the August system was developed the 3D animation capabilities of PCs were not good enough to render a high-quality lip-synchronized talking head[2]. To be able to get a responsive and smoothly animated August system the talking head module had to be run on a separate computer with better animation capabilities, while the rest of the dialogue modules were run on a PC with Linux. To simplify the implementation of this distributed system, a Broker architecture for communication between clients and servers was developed (Lewin 1998). In AdApt, all messages between modules were encoded in XML, in order to make it easier to handle multimodal messages (Beskow et al. Forthcoming).

These and other architecture related decisions are listed in Table 2. The table is also supposed to show the new demands each system had on the system architecture, in addition to the demands of earlier systems.

[1] According to Moore's Law the speed and memory are doubled every eighteen months (Moore 1965, 1997). In 1992 a PC would run at a speed of about 30 MHz with a memory of 16 Mb, while PCs in 2002 run at a speed of 2.8 GHz with a possibility of 1.5 Gb of memory.

[2] In 1998 a 3D graphics card on a PC could produce 10 frames per second, compared to 100 frames per second in 3D cards in 2002.

Table 2. *A list of architecture decisions in the different dialogue systems.*

System	System Requirements	Added system architecture feature
Waxholm	Multiple modules: synthesis, ASR, Parser, Dialogue Manager (DM), GUI, as well as a group of developers	Modularized architecture, which led to the development of a message handling system. Each module could use this system as a C-library that could be linked in during compilation.
GULAN	Developers at multiple locations	The dialogue modules were provided with Tcl-wrappers, which made them easier to use. The application and its GUI were built using Tcl/Tk.
	Users at multiple locations	By using the Tcl-plugin to Netscape it was possible to build a web-based version of Gulan, which students could use over the Internet.
AUGUST	System on multiple computers	A distributed architecture was developed, with a central Broker that handled all the messages that registered clients and servers wanted to send to each other.
	System in public environment	All servers that registered with the Broker had to provide information on how to restart them if they did not respond.
	System with multiple domains	An Input Manager and a domain predictor were used, making it possible to have several different dialogue managers for the different domains.
ADAPT	A system with multimodal output as well as multimodal input.	An Input/Output Manager was used, which merged input from different modalities into one message to the DM and decomposed a multimodal output message to the different output modules. Messages were encoded in XML to simplify handling of multimodal commands.
	Multiple graphical output modules (Map, Agent and Icon handler) that needed to get access to parts of the same GUI.	A GUI manager was developed which provided a common window with frames that it lent to the different output modules.
	The system used an open microphone, which led to a lot of fragmented user utterances.	The I/O Manager was responsible for sending only those utterances that were complete to the DM, incomplete utterances led to a facial expression of continued attention. It also handled timing in the system.
Pixie	Multiple simultaneous users, multiple user terminals in a public exhibition.	A central database was used where information about the users was stored. Users had to register (providing age and gender) before interacting with Pixie. The database also stored what they did at each terminal (a smart card was used to identify the users).
	The system was used in an exhibition that had thirty-minute shows.	The communication had to be asynchronous to make it possible for the exhibition manager to stop the interaction or for the users to go to the next terminal.

1992

1996

1998

2000

2002

4.3. System features

This thesis describes the development of spoken dialogue systems and subsequent empirical user studies. The collected dialogue corpora are influenced by a number of system features. This section tries to give an overview of some of the features of the corpus collection systems that might have influenced the human–computer interaction.

The behavior of the users is influenced by the type of domain, as well as the way in which the users acquired their interaction goals, see Table 3. Users of task-oriented dialogue systems have explicit goals with their interaction. In the simulated information retrieval dialogues of Waxholm and AdApt the goals were given in written or graphical scenarios. This made it possible to assess if the users had succeeded or not. In the user study with the fully automated AdApt system the users were only given the general goal of browsing the available apartments (Edlund et al. 2002). In the context-oriented dialogues with August and Pixie the users have no other goals than interacting with the system and get to know more about the agent or exhibition. This makes it hard to assess task success rates in these dialogues.

Table 3. *The domains of the systems and the origin of the users' goals.*

System	Domain	Domain type	User goals
Waxholm (woz)	Information Retrieval	Task-oriented	Written scenario
August (sys)	Persona at Exhibition	Context-oriented	Not any given
AdApt (woz)	Information Retrieval	Task-oriented	Pictorial scenario
AdApt (sys)	Information Browsing	Exploring	Find an apartment
Pixie (domain)	Home control	Task-oriented	Help Pixie
Pixie (social)	Guide at Museum	Context-oriented	Get to know Pixie

The systems' capabilities to understand also influenced the user interaction, and some of these features are listed in Table 4. If the system understands everything, users will behave differently than if it understands very little, since users tend to adapt their language to the understanding capabilities of the dialogue partner. The fully automated systems August and Pixie had limited understanding capabilities, which resulted in dialogues filled with misunderstandings. However, these corpora are interesting for two reasons: firstly they show what people would like to ask the system, secondly they show how people change their way of speaking to make the system understand them better when things go wrong. From a system development perspective a fully developed system gives experiences that are hard to gain in simulated settings. In the Waxholm simulations the understanding was limited by the domain specific parser and the dialogue manager, while the AdApt WOZ system simulated all parts of the input understanding.

However, the human wizard in the AdApt WOZ study tried to understand only utterances that were part of the domain and that were not too syntactically complex. Furthermore, the wizard was limited in the output generation by a fixed number of text templates.

Another relevant system feature is whether the user or the system controls the course of the dialogue. The system driven Pixie dialogues might for example be less interesting to study from a discourse perspective, than the mixed initiative AdApt dialogues.

Table 4. *Some system features of the corpus collection systems.*

System	Understanding restrictions	Initiative	Turn handling
Waxholm (woz)	Simulated ASR and a domain specific parser	Mixed	Push-to-talk, text button
August (sys)	500 words ASR bigram grammar semantic analyzer	User	Click-to-talk, facial expressions
AdApt (woz)	Simulated understanding and dialogue management	Mixed	Speech detection, facial expressions
AdApt (sys)	3000 words statistical ASR grammar and domain specific parser	Mixed	Speech detection, facial expressions
Pixie (domain)	Dynamic ASR grammar, average 280 states and 400 transitions	System	Speech detection, microphone icon
Pixie (social)	Static ASR grammar, 1500 states and 2000 transitions	User	Speech detection, microphone icon

Furthermore, the turn-handling strategies used in the systems are very important. In Waxholm and August the users had to push a button before speaking while AdApt and Pixie used speech detection, thus allowing the users to talk without having to push any buttons. This influenced the users' ways of speaking, resulting in a large number of fragmented utterances in the AdApt dialogues. The use of speech detection made turn-handling harder in the AdApt and Pixie systems. In these systems both facial expressions and graphical icons were used to simplify the turn-handling. None of the systems used barge-in, because an intelligent barge-in must be able to distinguish between turn-taking speech and backchanneling or self-directed speech. To train such a module the users' speech during system output must be recorded, transcribed and analyzed. However, for practical reasons the users' speech during system output was not recorded in any of the corpus collections, except for the user test on the fully automated AdApt system, which was videotaped. These recordings show that some users talked while the system was not listening. However, more empirical data is needed to be able to build an intelligent barge-in module for AdApt.

4.4. A description of the data collection

A number of dialogue corpora have been collected and transcribed. For each corpus collection a simulated or fully automated dialogue system has been developed, where the interactions have been logged and sound files have been saved. Transcription tools have been developed using the Snack sound toolkit (Sjölander 1997). Each of these has been designed to simplify the process of annotating certain features of the dialogue corpora and to make it easy to listen to certain parts of the dialogues. An overview of the collected and analyzed corpora is shown in Table 5.

Table 5. *An overview of all dialogue corpora.*

System	No. of users	No. of utterances	No. of words	No. of word types
Waxholm (WOZ)	68	1.912	10.829	694
August	2.500	10.058	41.330	2.968
Adapt (WOZ)	32	1.845	13.970	1.180
Adapt (SYS)	25	3.939	17.494	900
Pixie (domain)	1560	6.324	21.569	1.179
Pixie (social)	1346	11.259	22.480	1.491

As can be seen it has either been long and controlled dialogue collections from a few subjects at the lab, or short and uncontrolled interactions from thousands of users in public settings. Some features of the settings of the data collection are shown in Table 6. Who the users are and why they interact with the system might also be of interest. The Waxholm and AdApt dialogue collections were conducted at KTH, where students, friends and colleagues got a small reward for participating. The August and Pixie systems were exhibited in publicly available places where anybody could talk to them. Hence, the users of these systems are more representative of the general public and the settings of the data collection were more realistic than controlled laboratory settings. A problem with the August data is to ascertain who the users were. This was solved in the Pixie system by encouraging the users to register before they could interact with the system. In both these systems it was hard to know what the users' goals were and it was not possible to interview the users afterwards as was done when collecting the Waxholm and AdApt data.

Table 6. *Some facts about how the data was collected.*

System	Place	Users/Subjects	Reward	System Introduction
Waxholm(woz)	Silent room at KTH	students and colleagues	T-shirt	Verbal from Wizard
August	Exhibition public place	visitors general public	None	None
AdApt (woz)	Office at KTH	friends and colleagues	Cinema tickets	Verbal from experiment leader
AdApt (sys)	Office at KTH	friends and students	Cinema tickets	Minimal
Pixie (domain)	Museum	visitors general public	Users paid	Movie, museum guide
Pixie (social)	Museum	visitors general public	Users paid	Movie, museum guide

In the Waxholm and AdApt simulations, the experimental leader gave the subjects a short introduction to the system, and gave them the written scenarios to be regarded as hints on what to look for. Further to this, the experimental leaders told the subjects that they could interact for as long as they liked and that they could say “thank you” or “goodbye” when they wanted to finish. However, since the users of the AdApt system browsed the available apartments it was hard for them to know when to finish, which meant that the experiment leader had to enter the room after some time and ask the subject to finish.

There were no experimental leaders or guides present that could introduce the users of the August system to the system or its domains. It was placed in an exhibition space with concurrent exhibitions, but these were all unrelated to August. The users would often start the interaction by asking personal questions about August. Bell and Gustafson (1999a) investigated whether it was possible to make the users talk about the different domains instead of merely socializing. Some of the system’s utterances functioned as suggestions for possible topics of conversation. In some cases these were triggered by mistakes due to recognition errors. The effects of these mistakes were studied to see if the users followed these hints even though they had been socializing with the system before they occurred. The study showed that users often followed these hints and started asking domain related questions.

Pixie, on the other hand was an integrated part of the future home in the ‘Tänk Om’-exhibition. The users were introduced to Pixie in a short movie that all visitors saw before entering the exhibition. Furthermore, there were guides present that told the users to how to interact with Pixie and that asked them to help Pixie by talking to her. After engaging in a number of system driven dialogues, Pixie encouraged the visitors to ask questions about herself or about the exhibition. The visitors often followed these instructions, and asked questions about Pixie's personal appearance or interests, or questions about things they could see in the exhibition.

Users of the August system found it difficult to know what to say, which led to a large number of out-of-domain questions like *What is the capital of Finland?* and uncooperative dialogues in which the users tested the limits of the system by asking questions like *What is my name?* The same type of problematic utterances were also found in the Pixie dialogues with user initiative, where both out-of-domain utterances like *What is fifteen times twenty?* and uncooperative questions like *How long is a train?* were found. However, as can be seen in Figure 20, these two categories of problematic utterances were much more common in the August corpus than in the Pixie corpus.

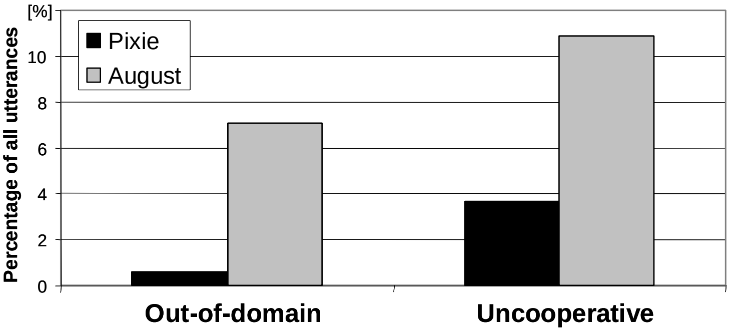


Figure 20. *The number of problematic utterances in the August corpus and in the Pixie corpus of dialogues where the users had the initiative.*

It seems like the users of the Pixie system found it easier to understand which the domains of the system were and they were less inclined to try to trick the system. This can be explained by the fact that Pixie and her role in the apartment was introduced in a movie, and that the museum guides explained how the users should interact with her. The users did not have to come up with something to say out of the blue, but could engage in dialogues about Pixie's abilities to change things in the apartment or to ask her to inform them about items that could be found there. Such context-oriented dialogues are possible to handle if the embodied character is given a natural role in its environment, and if users are introduced to the kinds of tasks it can handle.

This concludes the overview of the five spoken dialogue systems and some of the findings in the analyses of the collected dialogue corpora. The next chapter will give a short introduction to the included papers, and some examples of the findings they have reported on.

Chapter 5

OVERVIEW OF THE INCLUDED PAPERS

Paper I

Spoken dialogue data collection in the Waxholm project

Bertenstam, J. Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa–Leitao, A., Nord, L. and Ström, N.
STL-QPSR 1/1995

This paper describes the spoken dialogue system Waxholm, the WOZ experiments conducted with this system, the corpus of computer directed dialogues collected in these experiments and analyses of these data. The analyses include an extensive account of observed dialogue phenomena, some basic linguistic analyses of the user utterances, as well as an analysis of what the users said in different dialogue contexts. The paper also reports on the performance of the system.

The author of this thesis was one of the two Wizards in the WOZ simulations, and was involved in the dialogue design of the system, using the graphical interface for the STINA dialogue manager developed by Rolf Carlson (Carlson 1996). The author of this thesis performed the dialogue analysis presented in the paper.

Paper II

How do System Questions Influence Lexical Choices in User Answers?

Gustafson, J., Larsson, A., Carlson, R. and Hellman, K.
Eurospeech 1997

This paper describes the WOZ experiments performed with a unimodal spoken dialogue system in the travel domain and computerized questionnaire about travel plans. The experiments investigated how the systems verbal choices influenced the users' verbal input. Users were found to reuse the terms used by the system when answering system questions. The experiments showed that people adapt their answers to the system questions, by reusing the vocabulary and syntactic structures.

The author of this thesis designed the WOZ simulation environment and analyzed the collected data.

Paper III

Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech

Bell, L. and Gustafson, J.

ICPhS 1999

This paper contains a more detailed analysis of the strategies people used when August had failed to handle their previous utterance. The paper focuses on the phonetic realization of repeated utterances that had exactly the same words as the original utterances. Features such as hyperarticulation, inserted pauses and slower speaking rate were often found in the repeated utterances. Adults tended to shift the primary focus of the repeated utterances, while children often talked louder.

The author of this thesis developed the main parts of the August system, and collaborated with the co-author on the data analysis and corpus studies.

Paper IV

Speech Technology on Trial: Experiences from the August System

Gustafson, J. and Bell, L.

Journal of Natural Language Engineering 2000

This paper describes the multimodal spoken dialogue system August. Furthermore, the August dialogue corpus is described and analyzed. In particular, the paper discusses how the system design influenced the users' behavior and how the users behaved during error resolution. Both identical and non-identical repetitive sequences were analyzed, examining phonetic, lexical and syntactic aspects of linguistic adaptation. In the non-identical repetitions the most common change was the exchange of one lexical item for another. When an utterance was repeated users tended to alternate a specific feature, e.g. increased/decreased syntactic complexity. The paper also presents an analysis of what the users talked about with August. The corpus was tagged with utterance types that were supposed to reflect the users' intentions. In the August corpus about 40% of the utterances were tagged as 'socializing'.

The author of this thesis developed the main parts of the August system, and collaborated with the co-author on the data analysis and corpus studies.

Paper V

Modality Convergence in a Multimodal Dialogue System

Bell, L., Boye, J., Gustafson, J., and Wirén, M.
Gotalog 2000

This paper describes the multimodal spoken dialogue system AdApt and the design of a WOZ experiment that was aimed at studying the users' multimodal behavior. The aim was to investigate how the system's choice of reference influenced the users' verbal and graphical input to the system. Two versions of the system were used, one which used verbal only references and one that used multimodal references by shaking an icon while saying *This one has a balcony*. The study showed that the subjects were clearly influenced by the verbal references but less by the graphical. However, several users said in the post-experiment interviews that they did not quite understand what actions they were able to perform using mouse input.

The author of this thesis both designed the WOZ experiments and implemented the WOZ system in collaboration with other members of the AdApt group. He was also responsible for the data collection together with the first author.

Paper VI

Positive and Negative User Feedback in a Spoken Dialogue Corpus

Bell, L., and Gustafson, J.
ICSLP 2000

This paper examines feedback strategies in the AdApt WOZ corpus. The aim of the study was to investigate how users express positive and negative feedback to a dialogue system and to discuss the function of these utterances in the dialogues. User feedback in the AdApt corpus was labeled and analyzed, and its distribution in the dialogues is discussed. In the AdApt WOZ corpus 18% of all utterances were labeled as containing feedback. Feedback was often used to comment on the system's answer before asking the next question. Feedback was rarely provided in a turn of its own, but sometimes there was a short pause between the feedback and the next question. Only one subject avoided feedback altogether. The paper also discusses whether it is possible to utilize user feedback in the process of identifying errors in spoken dialogue systems.

The author of this thesis both designed the WOZ experiments and implemented the WOZ system in collaboration with other members of the AdApt group. The author collaborated with the co-author on the data analysis and corpus studies.

Paper VII

A Comparison of Disfluency Distribution in a Unimodal and a Multimodal Speech Interface

Bell, L., Eklund, R. and Gustafson, J.
ICSLP 2000

This paper describes an analysis of disfluency rates in the AdApt WOZ corpus of multimodal interactions and in the Telia Research Travel system corpus of unimodal interactions. The aim of the paper is to analyze and discuss the effects of modality, task and interface design on the distribution and frequency of disfluencies in these two corpora. The unimodal corpus was found to contain more disfluencies. A reason for this might be that the unimodal system used a very open greeting utterance and allowed the users to say very long and complex utterances. This made the user utterances longer during the entire unimodal dialogues, which partly explains the higher number of disfluencies in this corpus.

The author of this thesis both designed the multimodal WOZ experiments and implemented that WOZ system in collaboration with other members of the AdApt group. He performed the dialogue analysis of both the travel and the AdApt corpora together with the first author.

Paper VIII

Real-time Handling of Fragmented Utterances

Bell, L., Boye, J. and Gustafson, J.
NAACL workshop on Adaptation in Dialogue Systems 2001

This paper investigates the problem of turn-handling in the AdApt system. The AdApt WOZ corpus contained a large number of fragmented user utterances, and it was difficult for the system to decide when an utterance had ended. The paper proposes a method of using syntax and discourse to handle fragmented utterances and describes how it was implemented in the AdApt System.

The author of this thesis both designed the WOZ experiments and implemented the WOZ system in collaboration with other members of the AdApt group. He performed the analysis of the fragmented utterances together with the first author. He was responsible for adding the I/O handler that would handle fragmented utterances. This was incorporated into the AdApt system in collaboration with the second author.

Paper IX

Constraint Manipulation and Visualization in a Multimodal Dialogue System

Gustafson, J., Bell, L., Boye, J., Edlund, J. and Wirén, M.
ISCA workshop on Multimodal Interfaces in Mobile Environments 2002

This paper describes how the fully implemented AdApt system was developed and reports on some findings in the early user studies. It focuses especially on the problems the users had in understanding and influencing the way their input was processed by the system. The paper suggests that the search constraints that have been found in a user utterance would be displayed as graphical icons instead of given as verbal feedback. This could make the interaction faster and make it easier for the user to detect and correct misunderstandings. It also discusses the problem of constraint relaxation, and proposes that if automatic constraint relaxation is to be used, it would be helpful to visualize these as graphical icons as well, making it possible to show all constraints that the system is using for the moment. This would make browsing the information easier, since the user gets an overview of all constraints and has the possibility to use multimodal input to change a particular constraint.

The author of this thesis came up with the idea of adding the icon handler that handled the visualization of the constraints, and collaborated with the third and fourth authors to implement and incorporate it into the AdApt system

Paper X

Voice Transformations for Improving Children's Speech Recognition In A Publicly Available Dialogue System

Gustafson, J. and Sjölander, K.
ICSLP 2002

This paper describes the "Tänk Om" ("What If") exhibition and the Pixie dialogue system. It discusses the problem of collecting spoken dialogues with children without having a speech recognizer which is trained on children's speech. It is important to collect computer directed dialogues from children to be able to develop acoustic speech recognition for future systems. The paper describes a method of improving the recognition rates for children when using a commercial speech recognizer which is trained on adult speech and that uses telephone bandwidth. The children's speech is transformed on-the-fly before being down-sampled to telephone bandwidth and then sent to the speech recognizer. Two transformation methods were tested, one inspired by the Phase Vocoder algorithm and another by the Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) algorithm. Recognition errors could be reduced by something in the order of 30 to 45 percent if children's voices were transformed before the signal was down-sampled to telephone bandwidth.

The author of this thesis came up with the idea of transforming the children's voices on-the-fly before sending it to the recognizer. He also collaborated with the second author on incorporating the transformation software, developed by the second author, into the Pixie system.

Chapter 6

CONCLUDING REMARKS

This thesis has described the development of a number of multimodal spoken dialogue systems. The work has aimed at acquiring an understanding of how to build spoken dialogue systems that allow users to interact naturally. The need for data collection and the importance of analyzing dialogue corpora has been stressed. The collected dialogue corpora in this work have been used to train Swedish acoustic models for the Waxholm system, statistical recognition grammars for the fully automated AdApt system and as inspiration and data for building parsers and dialogue managers for both the Waxholm and the AdApt systems.

The experiences from the Waxholm and AdApt dialogue collections show the difficulty in collecting dialogue data from subjects who are given artificial goals. These data also show that users are influenced by the system output in the dialogues. However, this user behavior might be used to implicitly influence the users to say things that the system is able to understand. Other aspects of the system design also influenced the interaction, e.g. an open microphone in combination with a graphical user interface led to fragmented user utterances.

While developing the systems a number of problems had to be solved: Fragmented utterances which led to the development of a input/output handler in AdApt; turn-taking problems which led to the introduction of turn-taking gestures in August and AdApt, a microphone icon in Pixie and constraint icons in AdApt; problems of recognizing children's speech using a recognizer trained on adults led to the incorporation of a voice transformation module in the Pixie system.

The analyses of the speech corpora have given some insight into how people adjust their way of talking to computers. For example, users tend to use simple syntax, a quite small lexical variation and unambiguous pronouns. Furthermore, if the system fails they move towards a hyperarticulate pronunciation, sometimes inserting pauses between the words.

The experiences from the August and Pixie system show how naïve users would interact with an animated agent in a public setting. They indicate that people socialize if they are not given a specific task, but that these dialogues often are context-oriented, which means that they are possible to handle if the system set-up is controlled. Furthermore, the users of these publicly available systems were cooperative most of the time and willing to talk about the topics of conversation the animated agent suggested.

Finally, it was stated that context-oriented dialogues will be important in speech-enabled computer games. The author of this thesis will explore this issue in the three year EU project NICE (www.niceproject.com). NICE aims at developing 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 users will be able to refer multimodally to objects in their shared spatial context. It will be possible to generate context-oriented dialogues, since the system will know what is shown in a particular scene and since the personality and traits of the characters will be indicated by their appearance, movements and speech.

LIST OF PUBLICATIONS

The following papers have been published by the author. Articles marked with "●"are included in this thesis, those marked with "○" are not.

● Gustafson, J. and Sjölander, K. (2002) "Voice Transformations For Improving Children's Speech Recognition In A Publicly Available Dialogue System". *Proceedings of ICSLP02*, Colorado USA, vol. 1, pp 297–300.

● 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 Multi-Modal 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.

○ Lindberg, N. & Gustafson, J. (2000) "Example based shallow semantic analysis in the August spoken dialogue system". *STL-QPSR 1/00*, pp. 39–44.

● 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, pp. 29–34.

● Gustafson, J. & Bell, L. (2000) "Speech Technology on Trial: Experiences from the August System". *Journal of Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, pp. 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*, vol. 2 pp. 134–137.

● Bell, L. & Gustafson, J. (2000) "Positive and Negative User Feedback in a Spoken Dialogue Corpus". *Proceedings of ICSLP 00*, vol. 1, pp. 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*, vol. 3, pp. 626–629.

○ Gustafson, J., Sjölander, K., Beskow, J., Granström, B. & Carlson, R. (1999) "Creating web-based exercises for spoken language technology". *Proceedings of IDS '99*, tutoriaÎ session, pp. 165–168.

○ Gustafson, J., Lundeberg, M. & Liljencrants, J. (1999) "Experiences from the development of August – a multimodal spoken dialogue system". *Proceedings of IDS '99*, pp. 61–64.

○ Bell, L. & Gustafson, J. (1999a) “Utterance types in the August System”. *Proceedings of IDS ’99*, pp. 81–84.

○ Bell, L. and Gustafson, J. (1999b) “Interaction with an animated agent in a spoken dialogue system”. *Proceedings of Eurospeech ’99*, vol. 3, pp. 1143–1146.

● Bell, L. and Gustafson, J. (1999c) “Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech”. *Proceedings of ICPhS ’99,* vol. 2 pp. 1221–1224.

○ Gustafson, J., Lindberg, N. & Lundeberg, M. (1999) “The August spoken dialogue system”. *Proceedings of Eurospeech ’99*, vol. 3, pp. 1151–1154.

○ Gustafson, J., Elmberg, P., Carlson R. & Jönsson, A. (1998) “An Educational Dialogue System With a User Controllable Dialogue Manager”. *Proceedings of ICSLP ’98*.

○ Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., and Granström, B. (1998) “Web-based Educational Tools for Speech Technology”. *Proceedings of ICSLP ’98*.

○ Carlson, R., Granström, B., Gustafson, J., Levin, E. and Sjölander, K. (1998) “Hands-on Speech Technology on the Web”. Proceedings of the workshop ELSNET in Wonderland.

● Gustafson, J., Larsson, A., Carlson, R. & Hellman, K. (1997) “How do System Questions Influence Lexical Choices in User Answers?”. *Proceedings of Eurospeech ’97*, 22–25 September, Rhodes, Greece, vol. 4, pp. 2275–2278.

○ Sjölander, K. & Gustafson, J. (1997) “An Integrated System for Teaching Spoken Dialogue Systems Technology”. *Proceedings of Eurospeech ’97*, 22–25 September, Rhodes, Greece, vol. 4, pp. 1927–1930.

Ω Gustafson, J. (1996) *A Swedish Name Pronunciation System*, Licenciate Thesis, the Department of Speech, Music and Hearing, KTH, Stockholm.

○ Gustafson, J. (1995) “Using Two-level Morphology To Transcribe Swedish Names “, *Proceedings of Eurospeech ’95*, Madrid, Spain.

○ Bertenstam, J. Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa–Leitao, A., Nord, L. and Ström, N. (1995) “The Waxholm Application Data-Base”. *Proceedings of Eurospeech ’95*, vol. 1, pp. 833–836, Madrid, Spain.

○ Gustafson, J. (1995) “Transcribing names with foreign origin in the ONOMASTICA project”. *Proceedings of ICPhS ’95*, Stockholm, Sweden, August 13–19.

○ Carlson, R., Hunnicutt, S. and Gustafson, J. (1995) “Dialogue management in the Waxholm system”. *Proceedings of the Workshop Spoken Dialogue Systems*, Vigsø, Denmark.

○ Bertenstam, J. Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa–Leita, A., Nord, L. and Ström, N. (1995) "The Waxholm system - a progress report". *Proceedings of the Workshop Spoken Dialogue Systems*, Vigsø, Denmark.

● Bertenstam, J. Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa–Leitao, A., Nord, L. and Ström, N. (1995) "Spoken dialogue data collection in the Waxholm project". *STL-QPSR* 1/1995, pp. 50–73.

○ Gustafson, J. (1994) "ONOMASTICA - Creating a multi-lingual dictionary of European names". *Proceedings of the 8th Swedish Phonetics Conference, FONETIK '94*, May 24–26, Lund, Sweden, pp. 66–69.

○ Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. (1993) "An experimental dialogue system: WAXHOLM". *Proceedings of Eurospeech '93*, Berlin, Germany, vol. 3, pp. 1867–1870.

○ Gustafson J. (1992) "Databashantering som del av ett talförståelsesystem (database handling as part of a speech understanding system)". Master of Science Thesis, Department of Speech Communication and Music Acoustics, KTH, Stockholm.

REFERENCES

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.–E. & Öhman, T. (1998) “Synthetic faces as a lipreading support”. *Proceedings of ICSLP ’98*, Sydney, Australia, 30 November – 4 December 1998.

Allen, J. (1997) “Tutorial: Dialogue modelling for spoken language systems”. *ACL/EACL workshop for “Interactive spoken dialogue systems: bringing speech and NLP together in real applications”*. July 1997.

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L. & Stent, A. (2001) “Towards conversational human–computer interaction”. *AI Magazine*, vol. 22(4).

Allen, J., Schubert, L., Ferguson, G., Heeman, P., Hwang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M. & Traum, D. (1995) “The TRAINS Project: A Case Study in Defining a Conversational Planning Agent”. *Journal of Experimental and Technical Artificial Intelligence*, vol. 7, no. 1, pp. 7–48.

Allen, J. & Perrault, C. (1980) “Analyzing intention in utterances”. *Artificial Intelligence* 15(3), pp. 143–178.

Allwood, J. (1976) “Linguistic Communication as Action and Cooperation”. *Gothenburg Monographs in Linguistics 2*, Göteborg University, Department of Linguistics.

Ammicht E., Gorin A. & Alonso T. (1999) “Knowledge Collection for Natural Language Spoken Dialog Systems”. *Proceedings of Eurospeech ’99*, Budapest, Hungary, 5–9 September 1999, vol. 3, pp. 1375–1378.

André, E., Klesen, M., Gebhard, P., Allen, S. & Rist, T. (1999) “Integrating Models of Personality and Emotions into Lifelike Characters”. *Proceedings of the Workshop on Affect in Interactions – Towards a new Generation of Interfaces in conjunction with the 3rd i3 Annual Conference*, Siena, Italy, October 1999, pp. 136–149.

Arai, K., J. Wright, G. Riccardi & Gorin, A. (1998) “Grammar Fragment Acquisition using Syntactic and Semantic Clustering”. *Proceedings of ICSLP ’98*, Sydney, Australia, December 1998.

Aust, H., Oerder, M., Seide, F. & Stenbiss, V. (1995) “The Philips automatic train timetable information system” *Speech Communication* 17, pp. 249–262.

Austin, J. (1962) *How to Do Things with Words*, Oxford: Oxford University Press.

Balentine, B. (1999) “Re-engineering the speech menu: A ‘device’ approach to interactive list-selection” In: Gardner–Bonneau, D. (ed.), *Human Factors and Voice Interactive Systems,* Boston, Kluwer Academic Publ., pp. 37–61.

Balentine, B. & Morgan, D. (1999) *How to build a speech recognition application*. Enterprise Integration Group, San Ramon, USA.

Bangalore, S. & Johnston, M. (2000) “Integrating Multimodal Language Processing with Speech Recognition”. Proceedings of ICSLP ’00, Beijing, China, 16–20 November 2000, vol. 2, pp. 126–129.

Baum, L. F. (1900) *The Wonderful Wizard of Oz*. New York, George M. Hill Co.

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, 4 June 2001.

Bell, L. & Gustafson, J. (1999a) “Utterance types in the August System”. *Proceedings of IDS ’99*, pp. 81–84.

Bell, L. & Gustafson, J. (1999b) “Interaction with an animated agent in a spoken dialogue system”. *Proceedings of Eurospeech ’99*.

Bell, L. & Gustafson, J. (1999c) “Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech”. *Proceedings of ICPhS ’99*, vol. 2, pp. 1221–1224.

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*, pp. 29–34.

Benoît C., Mohamadi T. & Kandel S. (1994) “Effects of Phonetic Context on Audio-Visual Intelligibility of French”. *Journal of Speech and Hearing Research* 37, pp. 1195–1203.

Benoît, C., Martin, J. C., Pelachaud, C., Schomaker, L. & Suhm, B. (2000) “Audio–Visual and Multimodal Speech Systems”. In: Gibbon, D. (ed.), *Handbook of Standards and Resources for Spoken Language Systems – Supplement Volume*. Kluwer.

Bernsen, N. O. (2002) “Multimodality in language and speech systems – from theory to design support tool”. Chapter to appear in Granström, B. (ed.), *Multimodality in Language and Speech Systems*, Dordrecht: Kluwer Academic Publishers.

Bernsen, N., Dybkjaer, H. & Dybkjaer, L. (1998) *Designing Interactive Speech Systems* Springer–Verlag, London.

Bertenstam, J. Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa–Leita, A., Nord, L. and Ström, N. (1995) “The Waxholm system – a progress report”. *Proceedings of the workshop Spoken Dialogue Systems*, Vigsø, Denmark.

Beskow, J. (1997) “Animation of Talking Agents”. *Proceedings of AVSP ’97*, September, Rhodes, Greece, pp. 149–152.

Beskow, J. (1995) “Rule-based visual speech synthesis”. *Proceedings of Eurospeech ’95*, 18-21 September, Madrid, Spain, vol. 2, pp. 299–302.

Beskow, J., Edlund, J. & Nordstrand, M. (Forthcoming) “A model for generalised multi-modal dialogue system output applied to an animated talking head”. In Minker, W., Bühler, D. and Dybkjær, L. (Eds) *Spoken Multimodal Human–Computer Dialogue in Mobile Environments*, Dordrecht, The Netherlands, Kluwer Academic Publishers. To be published in Spring 2003.

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K–E & Öhman, T. (1997) “The Teleface project – Multimodal Speech Communication for the Hearing Impaired”. *Proceedings of Eurospeech ’97*, Rhodes, Greece, September 1997.

Beskow, J. & McGlashan, S. (1997) “Olga – A Conversational Agent with Gestures”. *Proceedings of the IJCAI ’97 workshop on Animated Interface Agents – Making them Intelligent*, Nagoya, Japan, August 1997.

Beskow, J. Dahlquist, M., Granström, B., Lundeberg, M., Spens, K–E. & Öhman, T. (1997) “The Teleface project – Multimodal speech communication for the hearing impaired”. *Proceedings of Eurospeech ’97*, 22–25 September, Rhodes, Greece, vol. 4, pp. 2003–2006.

Bickmore, T. & Cassell, J. (2000) “‘How about this weather’ Social Dialog with Embodied Conversational Agents”. *Proceedings of the American Association for Artificial Intelligence (AAAI) Fall Symposium on “Narrative Intelligence*”. 3–5 November, Cape Cod, MA, pp. 4–8.

Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. (1993) “An experimental dialogue system: WAXHOLM”. *Proceedings of Eurospeech ’93*, Berlin, Germany, vol. 3, pp. 1867–1870.

Blomberg, M. & Elenius, K. (1978) “A phonetically based isolated word recognition system”. *Proceedings of the 96th meeting of the Acoustical Society of America*, Honolulu, 4–8 November 1978, *JASA*, supplement, No. 1, Fall 1978, p. 181.

Bohus, D. & Rudnicky, A. (2002) “LARRI: A Language-based Maintenance and Repair Assistant”. *Proceedings of the ISCA Workshop Multi-Modal Dialogue in Mobile Environments Kloster Irsee*, 17–21 June, 2002, Germany.

Bolt, R. (1980) “Put that there: Voice and gesture at the graphics interface”. *Computer Graphics*, vol. 14, no. 3, pp. 262–270.

Boyce, S. (1999) “Spoken natural language dialogue systems: User interface issues for the future”. In: Gardner–Bonneau, D. (ed.), *Human Factors and Voice Interactive Systems*, Kluwer Academic Publishers, Boston, pp. 205–235.

Boye, J. Wirén, M. Rayner, M. Lewin, I. Carter, D. & Becket, R. (1999) “Language-processing strategies and mixed-initiative dialogues”. *Proceedings of IJCAI ’99,* *Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 31 July–6 August, 1999, Stockholm, pp. 17–24.

Bregler, C., Covell, M. & Slaney, M. (1997) "Video rewrite: Driving visual speech with audio". *Proceedings of SIGGRAPH 1997*, ACM Press/ACM SIGGRAPH, Los Angeles, CA, Computer Graphics Proceedings, Annual Conference Series, ACM, pp. 353–360.

Brennan, S. (1996) "Lexical entrainment in spontaneous dialog". *Proceedings of ISSD*, Philadelphia, PA, pp. 41–44.

Brennan, S. (1990) "Conversation as Direct Manipulation". In: Laurel, B. & Mountford, S. J., (eds.), *The Art of Human–Computer Interface Design*, Addison–Wesley.

Brennan, S. E. & Clark, H. H. (1996) "Conceptual Pacts and Lexical Choice in Conversation". *Journal of Experimental Psychology: Learning, Memory and Cognition* 22(6), pp. 1482–1493.

Brooke, N. (1992) "Computer graphics synthesis of talking faces". In: G. Bailly, G., Benoît, C. & Sawallis, T. R. (eds.), *Talking machines: Theories, Models, and Designs*, North Holland, Elsevier, Amsterdam, pp. 505–522.

Brown, G. & Yule, G. (1983) *Discourse analysis*. Cambridge: Cambridge University Press.

Bull, M. (1996) "An analysis of between-speaker intervals". *Proceedings of the Edinburgh Linguistic department conference*, pp. 18–27.

Button, Graham (1990). "Going up a blind alley: Conflating conversation analysis and computational modelling". In: Luff, P., Gilbert, N. & Frolich, D. (eds.), *Computers and Conversation*, London, Academic Press, pp. 67–90.

Buxton, W. (1990) "The 'natural' language of interaction: A perspective on nonverbal dialogues". In: Laurel, B. (ed.), *The Art of Human–Computer Interaction*, Reading, MA., Addison–Wesley, pp. 405–416.

Carberry, S. (1990) *Plan Recognition in Natural Language Dialogue*, ACL-MIT Press Series on Natural Language Processing. Cambridge, MIT Press, 1990.

Carlson, L. (1983) *Dialogue games: An approach to discourse analysis*, Dordrecht, Holland: D. Reidel Publishing.

Carlson, R., Granström, B. & Hunnicutt, S. (1982) "A multi-language text-to-speech module". *Proceedings of ICASSP '82*, Paris, vol. 3, pp. 1604–1607.

Carlson, R. & Granström, B. (1976) "A text-to-speech system based entirely on rules," *Conference Rec. 1976 IEEE International Conference on ASSP*, Philadelphia, PA, pp. 686–688.

Carlson R. (1996). "The dialog component in the Waxholm system". *Proceedings of Twente Workshop on Language Technology. Dialogue Management in Natural Language Systems*, pp. 209–218.

Cassell, J. & Thorisson, K. (1999) "The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents". *Applied Artificial Intelligence* 13, pp. 519–538.

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C. & Vilhjálmsson, H. (2002) "MACK: Media lab Autonomous Conversational Kiosk". *Proceedings of Imagina '02*, February 12–15, Monte Carlo.

Cassell, J., Vilhjálmsson, H. & Bickmore, T. (2001) "BEAT: The Behavior Expression Animation Toolkit". *Proceedings of SIGGRAPH '01*, Los Angeles, CA, 2001, pp. 477–486.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjlmsson, H. & Yan. H. (1999) "Embodiment in conversational interfaces: Rea". *Proceedings of the Association for Computing Machinery (ACM) Special Interest Group on Computer–Human Interaction (SIGCHI)*, Pittsburgh, PA., May 1999, pp. 520–527.

Cassell, J., Torres, O. & Prevost, S. (1999) "Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation". In: Wilks, Y. (ed.), *Machine Conversations*, The Hague: Kluwer, pp. 143–154.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, T., Prevost, S. & Stone, M. (1994) "Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents". *Proceedings of ACM/SIGGRAPH '94*, Reading, MA., Addison–Wesley, pp. 413–420.

Chu–Carroll, J. (2000) "MIMIC: An adaptive mixed initiative spoken dialogue system for information queries". *Proceedings of the 6th ACL Conference on Applied Natural Language Processing*, Seattle, WA, May 2000, pp. 97–104.

Clark, H. (1994) "Managing problems in speaking". *Speech Communication* 15, pp. 243–250.

Clark, H. & Brennan, S. (1991) "Grounding in communication". In: Resnick, Levine, & Teasley, (eds.), *Perspectives on socially shared cognition*, Washington, DC: APA Books, pp. 127–149.

Clark, H. & Schaefer, E. (1989) "Contributing to discourse". *Cognitive Science* 13, pp. 259–294.

Clark, H. & Wilkes–Gibbs, D. (1986) "Referring as a collaborative process". *Cognition*, 22, pp. 1–39.

Cohen, P. (1992) "The role of natural language in a multimodal interface". *Proceedings of User Interface Software Technology (UIST '92) Conference,* Academic Press, Monterey, CA., pp. 143–149.

Cohen, P. & Oviatt, S. (1995) "The role of voice input for human–machine communication". *Proceedings of the National Academy of Sciences*, vol. 92(22), pp. 9921–9927.

Cohen, M. & Massaro, D. (1993) "Modeling coarticulation in synthetic visual speech". In: Thalmann, N. M. & Thalmann, D., (eds.), *Models and Techniques in Computer Animation*, Springer–Verlag, Tokyo, pp. 139–156.

Cohen, P. &. Levesque, H. (1990) "Performatives in a rationally based speech act theory". *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics,* Pittsburgh, pp. 79–88.

Cohen, P. & Perrault, R. (1979) "Elements of a plan-based theory of speech acts". *Cognitive Science* 3, pp. 177–212.

Colburn, A., Cohen, M. & Drucker, S. (2000) "The Role of Eye Gaze in Avatar Mediated Conversational Interfaces". *MSR-TR-2000-81*. Microsoft Research.

Cole, R., Massaro, D., de Villiers, J., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, P., Connors, P. Tarachow, A. & Solcher. D. (1999) "New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children". *Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, London, UK.

Coutaz, J., Nigay, L. & Salber, D. (1994) "Taxonomic Issues for Multimodal and Multimedia Interactive Systems". *Proceedings of ERCIM '94 workshop on Multimedia Multimodal User Interfaces*, Nancy, October 1994.

Dahlbäck, N. (1997) "Towards a Dialogue Taxonomy". In: Elisabeth Maier, Marion Mast, Susann LuperFoy (Eds.) *Dialogue Processing in Spoken Language Systems.* In: Springer Verlag Series LNAI-Lecture Notes in Artificial Intelligence 1236.

Dahlbäck, N. (1991) *Representations of Discourse – Cognitive and Computational Aspects*, Linköping Studies in Science and Technology 264, PhD Thesis, Linköping University.

Dahlbäck, N., Jönsson, A. & Ahrenberg, L. (1993) "Wizard of Oz Studies – Why and How". *Knowledge-Based Systems,* December 1993, vol. 6, no. 4, pp. 258–266.

Damper, R. (1984) "Voice-input aids for the physically handicapped". *International Journal of Man–Machine Studies* 21, pp. 541–553.

DeCarlo, D., Revilla, C., Stone, M. & Venditti, J. (2002) "Making discourse visible: Coding and animating conversational facial displays". *Computer Animation*, pp. 11–16.

De Carolis, B., Pelachaud, C., Poggi, I. de Rosis, F. (2001) "Behavior Planning for a Reflexive Agent". *Proceedings of IJCAI 2001*, Seattle, September 2001.

Dehn, D. & van Mulken, S. (2000) "The impact of animated interface agents: a review of empirical research". *International Journal of Human–Computer Studies*, vol. 52, pp. 1–22.

Donath, J. (2001) "Mediated Faces". In: Beynon, M., Nehaniv, C. L. & Dautenhahn, K. (eds.). Cognitive Technology: Instruments of Mind. *Proceedings of the 4th International Conference,* CI 2001, Warwick, UK, August 6–9**.**

Ducatel, K., Bogdanowicz, M., Scapolo, M., Leijten J. & Burgelman J.–C. (2001) *Scenarios for ambient intelligence in 2010.* An ISTAG report.

Duncan Jr., S. (1972) "Some signals and rules for taking speaking turns in conversations". *Journal of Personality and Social Psychology*, 23(2), pp. 283–292.

Dybkjær, L., Bernsen, N. O. & Dybkjær, H. (1996) "Grice incorporated. Cooperativity in spoken dialogue". *Proceedings of COLING '96*, The 16th International Conference on Computational Linguistics, Copenhagen, 5–9 Aug 1996, pp. 328–333.

Edlund, J. & Nordstrand, M. (2002) "Turn-Taking Gestures and Hourglasses in a Multi-Modal Dialogue System". *Proceedings of the ISCA Workshop Multi-Modal Dialogue in Mobile Environments,* Kloster Irsee, Germany.

Ekman, P. (1982) *Emotion in the human face.* New York: Cambridge University Press.

Ekman, P. (1979) "About brows: Emotional and conversational signals". In: Cranach, M. von, Foppa, K., Lepenies, W. & Ploog, D. (eds.), *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium,* Cambridge, Cambridge University Press, pp. 169–202.

Ekman, P., Friesen, W.V., Ellsworth, P. (1972) *Emotion in the Human Face: Guidelines for Research and an Integration of Findings.* New York, Pergamon Press Inc.

Ervin-Tripp, S. (1979) "Children's verbal turn-taking" In: Ochs & Schieffelin (eds.), *Developmental Pragmatics,* New York: Academic Press.

Ezzat, T. & Poggio, T. (1998) "MikeTalk: A Talking Facial Display Based on Morphing Visemes". *Proceedings of the Computer Animation Conference*, Philadelphia, PA, June 1998, pp. 96–102.

Fant, G. (1960) *Acoustic Theory of Speech Production.* The Hague, Netherlands, Mouton.

Fant, G. (1953) "Speech Communication Research". *IngenjörsVetenskapsAkademin* 24, Stockholm, Sweden, pp. 331–337.

Fisher, C. (1968) "Confusions among visually perceived consonants". *Journal of Speech and Hearing Research* 11, pp. 796–804.

Foner, L. (1997) "Entertaining Agents: A Sociological Case Study". In: Johnson, W. L. & Hayes–Roth, B. (eds.), *Proceedings of the First International Conference on*

Autonomous Agents (Agents '97), Marina del Rey, CA, ACM Press. INFSYS RR 1843-01-02 25, pp. 122–129.

Fraser, N. & Gilbert, G. (1991) "Simulating Speech Systems". *Computer Speech and Language* 5(1), pp. 81–99.

Furnas, G., Landauer, T., Gomez, L. & Dumais, S. (1987) "The vocabulary problem in human–system communication". *Communications of the ACM*, vol. 30(11), November 1987, pp. 964–971.

Gardner–Bonneau, D. (1999) "Guidelines for speech-enabled IVR application design". In*:* Gardner–Bonneau, D. (ed.), *Human Factors and Voice Interactive Systems*, Boston, Kluwer Academic Publishers, pp. 147–162.

Gaasterland, T., Godfrey, P. & Minker, J. (1992) "An Overview of Cooperative Answering". *Journal of Intelligent Information Systems*, vol. 1, pp. 123–157.

Good, C. (1979) "Language as social activity: Negotiating conversation". *Journal of Pragmatics* 3, pp. 151–167.

Goodwin, C. (1981) *Conversational Organization: interaction between speakers and hearers*, New York/London, Academic Press.

Gorin, A., Riccardi, G. & Wright, J. (1997) "How may I help you?". *Speech Communication* 23, pp. 113–127.

Grasso, M. (1997) *Speech Input in Multimodal Environments: Effects of Perceptual Structure on Speed, Accuracy, and Acceptance*. PhD Thesis, University of Maryland.

Grice, H. P. (1975) "Logic and conversation". In: Cole, P. & Morgan, J. L. (eds.), *Syntax and Semantics, vol. 9: Pragmatics*, New York, Academic Press, pp. 113–128.

Grosz, B. & Sidner, C. (1986) "Attention, Intention, and the Structure of Discourse". *Computational Linguistics* 12(3), pp. 175–204.

Guindon, R. (1988) "A multidisciplinary perspective on dialogue structure in user–advisor dialogues". In: Guindon, R. (ed.), *Cognitive Science and its Applications for Human–Computer Interaction*. Hillsdale, N.J, Erlbaum.

Gustafson J. (1992) "Databashantering som del av ett talförståelsesystem (database handling as part of a speech understanding system)". Master of Science Thesis, Department of Speech Communication and Music Acoustics, KTH, Stockholm.

Gustafson, J., Elmberg, P., Carlson R. and Jönsson, A. (1998) "An Educational Dialogue System With a User Controllable Dialogue Manager". *Proceedings of ICSLP '98*.

Gustafson, J., Lundeberg., M & Liljencrants, J. (1999) "Experiences from the development of August – a multimodal spoken dialogue system". *Proceedings of IDS '99*.

Gustafson, J., Sjölander, K., Beskow, J., Granström, B. & Carlson, R. (1999) “Creating web-based exercises for spoken language technology”. tutoriaÎ session in *Proceedings of IDS ’99,* pp. 165–168.

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*, vol. 2, pp 134–137.

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 Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany.

Gustafson, J. & Bell, L. (2000) “Speech Technology on Trial: Experiences from the August System”. *Journal of Natural Language Engineering: Special issue on Best Practice in Spoken Dialogue Systems*, pp. 273–286.

Gustafson, J. & Sjölander, K. (2002) “Voice Transformations For Improving Children's Speech Recognition In A Publicly Available Dialogue System”. *Proceedings of ICSLP ’02, Colorado USA*, vol. 1, pp. 297–300.

Guye–Vuillieme, A., Capin, T., Pandzic, I., Magnenat–Thalmann, N. & Thalmann, D. (1999) “Non-verbal Communication Interface for Collaborative Virtual Environments”. *The Virtual Reality Journal*, Springer, vol. 4, pp. 49–59.

Guyomard, M. & Siroux, J. (1988) “Experimentation in the specification of an oral dialogue”. In: Niemann, H., Lang, M. & Sagerer, G. (eds.), *Recent Advances in Speech Understanding and Dialog Systems*, Berlin, Springer Verlag, vol. 46, pp. 497–501.

Heeman, P. & Hirst, G. (1995) “Collaborating on referring expressions”. *Computational Linguistics* 21(3), pp. 351–382.

Hemphill, C., Godfrey, J. & Doddington, G. (1990) “The ATIS Spoken Language Systems, Pilot Corpus”. *Proceedings of 3rd DARPA Workshop on Speech and Natural Language*, Hidden Valley, PA, June 1990, pp. 102–108.

Hirschberg, J., Litman, D. & Swerts, M. (2000) “Generalizing prosodic prediction of speech recognition errors”. *Proceedings of ICSLP ‘00*, Bejing, China, September 2000.

Hirschberg, J. & Pierrehumbert, J. (1986)“The Intonational Structuring of Discourse.” *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pp. 136–144.

Hirschman, L. (1992) “Multi-site data collection for a spoken language corpus”. *Proceedings of the DARPA Speech and Natural Language Workshop*, Harriman, NY, February 1992. Morgan Kaufmann Publishers, Inc., pp. 7–14.

Hulstijn, J. (2000) *Dialogue Models for Inquiry and Transaction* PhD Thesis, Universiteit Twente.

Ibrahim, A., Lundberg, J. & Johansson, J. (2001) "Speech Enhanced Remote Control for Media Terminal". *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 2001.

Jakobson, R., Fant, G. & Halle, M. (1952) "Preliminaries to Speech Analysis". *MIT Acoustic Laboratory, Technical Report 13.*

Johnson, W. Rickel, J. & Lester, J. (2000) "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments". *International Journal of Artificial Intelligence in Education* 11, pp. 47–78.

Julia L. & Cheyer A. (1998). "Cooperative Agents and Recognition Systems (CARS) for Drivers and Passengers". *Proceedings of OzCHI '98*, Adelaide, Australia, pp 32–38.

Julia L., Cheyer A., Dowding J., Bratt H., Gawron J.M., Bratt E. & Moore R. (1998). "How Natural Inputs Aid Interaction in Graphical Simulations? " *Proceedings of VSMM '98*, Gifu, Japan, pp 466–468.

Jurafsky, D., Shriberg, E., Fox, B. & Curl, T. (1998) "Lexical, Prosodic, and Syntactic Cues for Dialog Acts". *Proceedings of ACL/COLING '98 Workshop on Discourse Relations and Discourse Markers*, pp. 114–120.

Jönsson, A. (1997) "A model for habitable and efficient dialogue management for natural language interaction". *Natural Language Engineering*, Cambridge University Press, vol. 3(2/3), pp. 103–122.

Jönsson, A. (1996) "Natural language generation without intentions". *Proceedings of ECAI '96 Workshop Gaps and Bridges: New Directions in Planning and Natural Language Generation*, pp. 102–104.

Jönsson, A. (1993) *Dialogue management for natural language interfaces.* PhD Thesis, Linköping University, Department of Computer and Information Science.

Jönsson, A. & Dahlbäck, N. (2000) "Distilling dialogues – A method using natural dialogue corpora for dialogue systems development". *Proceedings of 6th Applied Natural Language Processing Conference*, Seattle, 2000, pp. 44–51.

Jönsson, A. & Dahlbäck, N. (1988) "Talking to a Computer is not Like Talking to Your Best Friend". *Proceedings of The first Scandinivian Conference on Artificial Intelligence*, Tromsø, Norway, 9–11 March, 1988.

Kahneman, D. (1973) *Attention and Effort.* Prentice–Hall, Englewood–Cliffs, New Jersey.

Karl, L., Pettey, M. & Shneiderman, B. (1993) "Speech-Activated versus Mouse-Activated Commands for Word Processing Applications: An Empirical Evaluation". *International Journal of Man–Machine Studies*, 39(4), pp. 667–687.

Kennedy, A., Wilkes, A., Elder, L. & Murray, W. (1988) "Dialogue with machines". *Cognition* 30, pp. 73–105.

Kellner, A., Rueber, B. & Seide, F. (1996) "A voice-controlled automatic switchboard and directory information system". *Proceedings of the IEEE Third Workshop on Interactive Voice Technology for Telecommunications Applications*, Basking Ridge, NJ, pp. 117–120.

Koda, T. & Maes, P. (1996) "Agents with faces: The effects of personification of agents" *Proceedings of Human–Computer Interaction*, London, pp. 239–245.

Lambert, L. (1993) *Recognizing Complex Discourse Acts: A Tripartite Plan-Based Model of Dialogue*. PhD Thesis, University of Delaware, Department of Computer Science.

Lamel, L., Rosset, S. & Gauvin, J.–L. (2000). "Considerations in the design and evaluation of spoken dialogue systems". *Proceedings of ICSLP '00*, 16–20 October, Beijing, China, vol. 4, pp. 5–8.

Lamel, L. Rosset, S. Gauvin, J. Bennacef, S. Garnier–Rizet M. & Prouts, B. (1998) "The LIMSI ARISE System". *Proceedings IEEE 4th Workshop Interactive Voice Technology for Telecom Applications*, Torino, Italy, pp. 209–214.

Laurel, B. (1990) "Interface agents : Metaphors with character". In: Laurel, B. (ed.), *The Art of Human–Computer Interface Design*, Addison–Wesley, pp. 355–365.

Leech, G., Myers, G. & Thomas J. (1995) *Spoken English on computer: transcription, mark-up and application*. London, Longman.

Leijten, M. & Van Waes, L. (2001). "The influence of voice recognition on the writing process: cognitive and stylistic effects of speech technology on writing business texts". *Proceedings of Human–Computer Interaction: INTERACT '01,* 9–13 July 2001, Tokyo, Japan, pp. 783–784.

Lester, J. Voerman, J. Towns, S. & Callaway, C. (1999) "Deictic Believability: Coordinated Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents". *Applied Artificial Intelligence*, 13(4–5), pp. 383–414.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B. & Bhogal, R. (1997) "The persona effect: Affective impact of animated pedagogical agents". *Proceedings on Human Factors in Computing Systems*, Atlanta, GA, pp. 359–366.

Lester, J. & Stone, B. (1997) "Increasing believability in animated pedagogical agents". In: Johnson, W. L. & Hayes–Roth, B. (eds.), *Proceedings of the First International Conference on Autonomous Agents*, February, Marina del Rey, CA, pp. 16–21.

Lewin, E. (1998) "The Broker Architecture". `http://www.speech.kth.se/broker/`

Lewis, J. & Parke, F. (1987) "Automated Lip-synch and Speech Synthesis for Character Animation". *Proceedings of Graphics Interface '87*, Toronto, pp. 143–147.

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

Life, A., Salter, I., Temem, J., Bernard, F., Rosset, S., Bennacef, S. & Lamel, L. (1996) "Data Collection for the MASK Kiosk: WOZ vs. Prototype System," *Proceedings of ICSLP '96*, 3–6 October, Philadelphia, PA, pp. 1672–1675.

Liljencrants, J. (1967) "The OVE III speech synthesizer". *STL-QPSR2-3/1967*, pp. 76–81.

Lindberg, N. & Gustafson, J (2000) "Example based shallow semantic analysis in the August spoken dialogue system". *STL-QPSR* 1/00.

Lindblom, B. (1963) "Spectrographic study of vowel reduction". *JASA* 35, pp. 1773–1781.

Liquid Media (2002) "Pixie". A speech enabled animated agent that is part of the 'Tänk Om' exhibition at the Telecom museum in Stockholm. `http://www.liquid.se`

Litman, D. & Allen, J. (1990) "Discourse processing and commonsense plans". In: R. Cohen, P. R. Morgan J. & Pollack, M. E. (eds.), *Intentions in Communication*, ch. 17, MIT Press, pp. 365–388.

Lundeberg, M. & Beskow, J. (1999) "Developing a 3D-agent for the August dialogue system". *Proceedings of AVSP '99*, Santa Cruz, CA.

Martin, J. C. (1998) "TYCOON: theoretical and software tools for multimodal interfaces". In: Lee, J. (ed.), *Intelligence and Multimodality in Multimedia interfaces*, AAAI Press.

Martin, J. C., Julia, L. & Cheyer, A. (1998) "A Theoretical Framework for Multimodal User Studies". *Proceedings of the Second International Conference on Cooperative Multimodal Communication, Theory and Applications (CMC'98)*, 28–30 January 1998, Tilburg, The Netherlands.

Martin, J.C., Grimard, S. & Alexandri, K. (2001) "On the annotation of the multimodal behaviour and computation of cooperation between modalities". *Proceedings of the workshop on 'Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents'*, May 29, 2001, Montreal pp. 1–7.

Martin. T. B. (1976) "Practical applications of voice input to machines". *Proceedings of the IEEE*, 64(4), April 1976, pp. 487–501.

Mauldin, M. (1994) "Chatterbots, TinyMUDs, and the Turing Test: Entering the Loebner Prize Competition". *Proceedings of Twelfth National Conference on Artificial Intelligence (AAAI-94),* August 1994, Seattle, WA, pp. 16–21.

McBreen, H. & Jack, M. (2001) "Evaluating Humanoid Synthetic Agents in E-Retail Applications". *IEEE Transactions on Systems, Man and Cybernetics, Special Issue: The Human in the Loop*, vol. 31(5), p. 394.

McRoy, S. & Hirst, G. (1995) "The repair of speech act misunderstandings by abductive inference". *Computational Linguistics* 21(4), pp. 435–478.

Melin H., Sandell A. & Ihse, M. (2001) "CTT-bank: A speech controlled telephone banking system – an initial evaluation". *TMH-QPSR*, KTH, 1, pp. 1–27.

Moore, G. (1965) "Cramming more components onto integrated circuits". *Electronics Magazine*, vol. 38, April 1965, pp. 114–117.

Moore, G. (1997) "An Update on Moores Law." Transcript of a Keynote speech at Intel Developer Forum, San Francisco, CA, September 30, 1997, available at `http://www.intel.com/pressroom/archive/speeches/gem93097.htm`

Moody, T. (1988) *The Effects of Restricted Vocabulary Size on Voice Interactive Discourse Structure*. PhD thesis, North Carolina State University.

Nakatani, C. & Hirschberg, J. (1994). "A corpus-based study of repair cues in spontaneous speech". *JASA* 95(3), pp. 1603–1616.

Nass, C., Isbister, K. & Lee, E.-J. (2000) "Truth is beauty: Researching conversational agents". In: Cassells, J., Sullivan, J., Prevost, S. & Churchill, E. (eds.), *Embodied conversational agents*, Cambridge, MA, MIT Press, pp. 374–402.

Nielsen, J. (1993) "Noncommand user interfaces". *Communications of the ACM*, vol. 36(4), April 1993, pp. 83–99. Revised version is available at: `http://www.useit.com/papers/noncommand.html`

Nintendo (1999) "Hey, you, Pikachu!/Pikachu Genki Dechu". A N64 game published by Nintendo. `http://www.heyyoupikachu.com`

Novick, D., Hansen, S. & Marshall, C. (1999) "Limiting factors of automated telephone dialogues". In: Gardner–Bonneau, D. (ed.), *Human Factors and Voice Interactive Systems*, Boston, Kluwer Academic Publishers, pp. 163–186.

Noyes, J. (2001) "Talking and writing – how natural in human–machine interaction?". *International Journal of Human–Computer Studies*, vol. 55, pp. 503–519.

Nye, J. (1982) "Human factors analysis of speech recognition systems". *Speech Technology*, 1, 2, pp. 50–57.

Oerder, M. & Aust, H. (1994) "A realtime prototype of an automatic inquiry system". *Proceedings of ICSLP '94*, 18–22 September, Yokohama, Japan, vol. 2, pp. 703–706.

Oviatt, S. (2000) "Talking To Thimble Jellies: Children's Conversational Speech with Animated Characters". In: Yuan, B., Huang, T. & X. Tang, X. (eds.), *Proceedings of ICSLP '00*, Beijing, China, vol. 3, pp. 877–880.

Oviatt, S. (1999) "Ten myths of multimodal interaction". *Communications of the ACM*, vol. 42, no. 11, November 1999, pp. 74–81.

Oviatt, S. (1995) "Predicting spoken disfluencies during human–computer interaction". *Computer Speech and Language* 9(1), pp. 19–35.

Oviatt, S. (1992) "Pen/voice: Complementary multimodal communication". *Proceedings of Speech Tech '92*, February, New York, pp. 238–241.

Oviatt, S. & Cohen, P. (2000) "Multimodal Interfaces That Process What Comes Naturally". *Communications of the ACM*, vol. 43(3), March 2000, pp. 45–53.

Oviatt, S., DeAngeli, A. & Kuhn, K. (1997) "Integration and synchronization of input modes during multimodal human–computer interaction". *Proceedings of Conference on Human Factors in Computing Systems: CHI '97*, New York, ACM Press, pp. 415–422.

Oviatt, S., MacEachern, M. & Levow, G. (1998) "Predicting hyperarticulate speech during human–computer error resolution". *Speech Communication* 24(2), pp. 1–23.

Oviatt, S. & VanGent R. (1996) "Error resolution during multimodal human–computer interaction". In: *ICSLP '96*, 3–6 October, Philadelphia, PA, vol. 1, pp. 204–207.

Pandzic, I., Ostermann, J. & Millen, D. (1999) "User evaluation: Synthetic faces for interactive services". *The Visual Computer*, vol. 15, Issue 7/8, 4 November 1999, pp. 330–340.

Parke, F. (1975) "A model for human faces that allows speech synchronized animation". *Journal of Computers and Graphics* 1(1), pp. 1–4.

Pavlovi, V., Berry, G. & Huang, T. (1998) "A Multimodal Human–Computer Interface for the Control of a Virtual Environment". *American Association for Artificial Intelligence*.

Pelachaud, C., Badler, N. & Steedman, M. (1996) "Generating Facial Expressions for Speech". *Cognitive Science* 20(1), pp. 1–46.

Pelachaud, C. & Prevost, S. (1994) "Sight and Sound: Generating Facial Expressions and Spoken Intonation from Context". *Proceedings of the second ESCA Workshop on Speech Synthesis*, New Paltz, NY, USA, September 1994, pp. 216–219.

Peckham, J. (1991) "Speech Understanding and Dialogue over the Telephone: An Overview of the ESPRIT SUNDIAL Project". *Proceedings of the DARPA Speech and Natural Language Workshop*, Pacific Grove, CA., pp. 14–27.

Peissner, M., Heidmann, F. & Ziegler, J. (2001) "Simulating recognition errors in speech user interface prototyping". In: Smith, M. J., Michael, J. et al. (eds)., Usability evaluation and interface design: *Proceedings of HCI International,* Mahwah, NJ, Lawrence Erlbaum, pp. 233–237.

Petrelli, D., De Angeli, A., Gerbino, W., Cassano G. (1997) “Referring in multimodal systems: The importance of user expertise and system features”. *Proceedings of ACL*, pp. 14–19.

Poggi, I. & Pelachaud, C. (2000) “Gaze and its meaning in animated faces”. In: McKevitt, P. (ed.), *Language, vision and music*, Amsterdam, John Benjamins.

Poggi, I. & Pelachaud, C. (1998) “Performative faces”. *Speech Communication* 26, pp. 5–21.

Polanyi, L. & Scha, R. (1984) “A syntactic approach to discourse semantics”. *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting of the Association for Computational Linguistics*, Stanford University, pp. 413–419.

Qvarfordt, P. & Jönsson, A. (1999) “Evaluating the Dialogue Component in the Gulan Educational System”. *Proceedings of Eurospeech ’99*, Budapest, Hungary, pp. 643–646.

Power, R. (1979) “The Organisation of Purposeful Dialogues”. *Linguistics*, vol. 17, pp. 107–152.

Rayner, M. Lewin, I. Gorrell, G & Boye, J. (2001) “Plug and Play Speech Understanding”. *Proceedings 2nd SIGdial Workshop on Discourse and Dialogue*, September 2001.

Reeves, B. & Nass, C. (1996) *The Media Equation: how people treat computers, televisions and new media like real people and places* Cambridge, Cambridge University Press.

Rich, C., Sidner, C. & Lesh, N. (2001) “COLLAGEN: Applying Collaborative Discourse Theory to Human–Computer Interaction”. *Artificial Intelligence Magazine* 22(4), pp. 15–25.

Richards, M. & Underwood, K. (1984) “Talking to Machines. How are People Naturally Inclined to Speak?”. In: Megaw, E. D. (ed.), *Contemporary Ergonomics*, Taylor & Francis, pp. 62–67.

Rimé, B. & Schiaratura, L. (1991) “Gesture and speech”. In: Feldman, R. S. & Rimé, B. (Eds.), *Fundamentals of nonverbal behavior,* Cambridge: Cambridge University Press, pp. 239–281.

Rist, T., André, E. & Müller, J. (1997) “Adding Animated Presentation Agents to the Interface”. In: Moore, J., Edmonds, E. & Puerta, A. (eds.). *Proceedings of the 1997 International Conference on Intelligent User Interfaces*, New York, ACM Press, pp. 79–86.

Sacks, H., Schegloff, E. & Jefferson G. (1974) “A simplest systematics for the organisation of turn-taking for conversation”. *Language* 50, pp. 696–735.

Salber, D. & Coutaz, J. (1993) “Applying the Wizard of Oz Technique to the Study of Multimodal Systems”. In: Bass, L., Gornostaev, J. & Unger, C. (eds.), *Human*

Computer Interaction, 3rd International Conference EWHCI '93, Springer Verlag, vol. 753, pp. 219–230.

Schegloff, E. (1968) "Sequencing in conversational openings". *American Anthropologist* 70, pp. 1075–1095.

Schenkein, J. (1978) *Studies in the organisation of conversational interaction.* New York, Academic Press.

Scherer, K. (1980) "The functions of nonverbal signs in conversation". In: Giles, H. & St. Clair, R. (eds.), *The Social and Physiological Contexts of Language*, Lawrence Erlbaum Associates, pp. 225–243.

Schneiderman, B. (2000) "The Limits of Speech Recognition". *Communications of the ACM*, 43(9), September 2000, pp. 63–65.

Schneiderman, B. (1997) "Direct Manipulation for Comprehensible, Predictable, and Controllable User Interfaces". *Proceedings of IUI '97***,** pp. 33–39.

SEGA (2000) "Seaman". a dreamcast computer game published by SEGA, `http://www.sega.com/games/dreamcast/post_dreamcastgame.jhtml?PRODID=194`

Searle, J. R. (1969) *Speech Acts: An Essay in the Philosophy of Language.* Cambridge, Cambridge University Press.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P. & Zue, V. (1998) "Galaxy-II: A reference architecture for conversational system development". *Proceedings of ICSLP '98*, Sydney, Australia, vol. 3, pp. 931–934.

Sjölander, K. (1997) "The SNACK Speech Visualization Module for Tcl/Tk". `http://www.speech.kth.se/snack/`

Sjölander, K. & Gustafson, J. (1997) "An Integrated System for Teaching Spoken Dialogue Systems Technology". *Proceedings of Eurospeech '97*, 22–25 September, Rhodes, Greece, vol. 4, pp. 1927–1930.

Sinclair, J. M. & Coulthard, R. M. (1975) *Towards an analysis of Discourse: The English used by teachers and pupils*, Oxford University Press.

Smith, R., Hipp, R. & Biermann, A. (1992) "A Dialog Control Algorithm and Its Performance". In: Bates, M. & Stock, O. (eds.): *Third Conference on Applied Natural Language Processing*, 31 March–3 April, 1992, pp. 9–16.

Stein, A., Gulla, J. & Thiel, U. (1997) "Making sense of user mouse clicks: Abductive reasoning and conversational dialogue modeling". In: Jameson, A., Paris, C. & Tasso, C., (eds.), User Modeling: *Proceedings of the Sixth International Conference, UM '97*. Vienna/Wien/New York, Springer, pp. 89–100.

Ström, N. (1997) *Automatic continuous speech recognition with rapid speaker adaptation for human/machine interaction,* PhD Thesis, Department of Speech, Music and Hearing, KTH, Stockholm.

Takeuchi, A. & Naito, T. (1992) "Situated facial displays: Towards social interaction". *International Conference on Computer Human Interaction (CHI)*, Denver, CO, pp. 450–455.

Takeuchi A. & Nagao, K. (1993) "Communicative facial displays as a new conversational modality". *ACM/IFIP INTERCHI '93*, Amsterdam.

Thórisson, K. (2002) "Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action". In: Granström, B., House, D., Karlsson, I. (eds.), *Multimodality in Language and Speech Systems*, Dordrecht, The Netherlands, Kluwer Academic Publishers, pp. 173–207.

Traum, D. & Allen, J. (1992) "A speech acts approach to grounding in conversation". *Proceedings of ICSLP '92*, 12–16 October, Banff, Canada, vol. 1, pp. 137–40.

Traum, D. & Heeman, P. (1997) "Utterance units in spoken dialogue". In: Maier, E. Mast, M. & Luperfoy, S. (eds.), *Dialogue Processing in Spoken Language Systems* ECAI-96 Workshop, Lecture Notes in Artificial Intelligence, Springer-Verlag, Heidelberg, pp. 125–140.

Traum, D. & Hinkelman, E. (1992) "Conversation acts in task-oriented spoken dialogue". *Computational Intelligence* 8(3), Special Issue on Non-literal language, pp. 575–599.

Traum, D. & Rickel, J. (2001) "Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds". *Proceedings of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, August 2001, Seattle, WA, pp. 766–773.

van Mulken, S., André, E. & Müller, J. (1998) "The Persona Effect: How Substantial is it?". *Proceedings of HCI '98*, Sheffield, pp. 53–66.

Virzi, R. & Huitema, J. (1997) "Telephone-based menus: Evidence that broader is better than deeper". *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, Santa Monica, pp. 315–319.

Wahlster, W., Reithinger, N. & Blocher, A. (2001) "SmartKom: Towards Multimodal Dialogues with Anthropomorphic Interface Agents". *Proceedings of MTI Status Conference*, 26–27 October 2001.

Walker, M., Litman, D., Kamm, C. & Abella, A. (1997) "PARADISE: A Framework for Evaluating Spoken Dialogue Agents". *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL '97)*, San Francisco, Morgan Kaufmann, pp. 271–280.

Walker, J., Sproull, L. & Subramani, R. (1994) "Using a human face in an interface". *Proceedings of the CHI '94 Conference Companion on Human Factors in Computing Systems*, Boston, MA, pp. 85–91.

Waters, K. (1987) "A muscle model for animating three-dimensional facial expression". *Computer Graphics* 22(4), pp. 17–24.

Weizenbaum, J. (1966) "ELIZA – A computer program for the study of natural language communications between men and machines". *Communications of the Association for Computing Machinery*, 9, pp. 36–45.

Whittaker, S A Stenton, P. (1988) "Cues and control in expert-client dialogues". *Proceedings ACL '88*, pp. 123–130.

Victor, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J. & Seneff, S. (1991) "Integration of speech recognition and natural language processing in the MIT Voyager system" *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 713–716.

Zue, V. & Glass, J. (2000) "Conversational interfaces: advances and challenges". *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, vol. 88(8), pp. 1166–1180.

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T. & Hetherington, L. (2000) "Jupiter: A telephone-based conversational interface for weather information". *IEEE Trans Acoustics, Speech and Signal Processing*, vol. 8(1), pp. 85–96.

Öhman, S. (1966) "Coarticulation in VCV utterances: Spectrographic measurements". *JASA* 39, pp. 151–168.

Öhman, T. (1999) "A visual input module used in the August spoken dialogue system". *QPSR* 1-2/99, pp. 39-44.

This page was intentionally left blank

EPILOGUE

In 1950, in an article published in the scientific journal Mind, British Mathematician Alan Turing asked the question, "Can machines think?" He proposed a test, now known as the Turing Test, in which machines—computers—could be judged and evaluated on their "human" ability to "think" by asking them a series of idiosyncratic questions that most people can answer. His prediction: within fifty years, a computer program would be capable of mimicking human thinking to such a degree that it would fool a human interrogator more than 50% of the time. Every year since 1990, the Loebner Prize ($100,000 and an 18k gold medal) has been offered to the first computer program that can pass the Turing Test.

What Turing did not consider—and what experts in artificial intelligence have ignored—is an even weightier question: "Can humans think?" Every year since 1999, the Neuman Prize (a medal) has been offered to the first human that can pass the Neuman Test, a test designed to determine whether or not humans have the ability to think.

The most recent Neuman Prize competition took place near my home where, as judge, I administered the Neuman Test in a controlled environment by communicating anonymously (I was speaking into a microphone hidden inside a clown's head) with an equally anonymous and (presumably) "human" respondent. Here is the complete text:

Test date: May 24, 1999
Time of day: 12:21 PM (PDT)

"HUMAN":	**JUDGE:**
Next, please.	Hello. How are you?
Can I take your order?	How are you?
What?	How do you feel?
I feel fine.	How does Fine feel?
What do you want?	I want some of that old time religion. Can you name one?
You want a burger?	How did you know I wanted a burger?
One burger. With cheese?	Cheese? What is cheese?
You want fries?	You tell me. Do I want fries?
Is that a yes?	A yes is a negative no, isn't it?
Small or large?	Which do you recommend?
You get more for the money with a large.	Then I'd like the small. Does that make sense?
Yes sir. Anything to drink?	If one were thirsty for knowledge, where would one go?
Coke, Sprite, or Root Beer.	Please write me a sonnet on the subject of the Forth Bridge.
What?	Add 34,957 to 70,764.
No drink, sir?	Where is Ypsilanti?
What?	Where is Ypsilanti?
Okay, one burger, small fries, no drink— Is that all you're having?	Is that all I can have?
Anything else?	What's your favorite black-and-white movie?
Huh?	Do you play chess?
I'm sorry, I —	I have my king at King1, and no other pieces. You have your king at King6 and a rook at Rook1. It's your move. What do you play? *(There is a sustained honking of car horns.)*
I can't hear you.	Never mind. How much?
105,621.	What?
The sum of 34,957 and 70,764.	Oh.
Your total is two ninety-nine. Pull forward, pay at the next window. Next, please."	

A satirical text written by the humorist Matt Neuman called "Can humans think?", which is available at http://www.mattneuman.com/think.htm. Used by kind permission of Matt Neuman.